The Routledge Handbook of Vocabulary Studies

Edited by Stuart Webb

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2020
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

and by Routledge
52 Vanderbilt Avenue, New York, NY 10017

Routledge is an imprint of the Taylor & Francis Group, an informa business

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book has been requested

ISBN: 978-1-138-73572-9 (hbk)
ISBN: 978-0-429-29158-6 (ebk)

Typeset in Times New Roman
by Apex CoVantage, LLC

29
Measuring Lexical Richness

Kristopher Kyle

Introduction

Indices of lexical richness are used to analyze characteristics of lexical use in written and spoken responses. The term lexical richness was introduced by Yule (1944) to refer to the number of words in a particular author's vocabulary (as evidenced by the words used in their published works). Although Yule had a particular method of calculation in mind, subsequent researchers have used the term to refer to a number of different constructs, such as the propositional density of the words in a text (lexical density; Halliday, 1985; Linnarud, 1986; Lu, 2012; Ure, 1971), the diversity or variation of words in a text (lexical diversity; Malvern, Richards, Chipere, & Durán, 2004; McCarthy & Jarvis, 2007), and/or the proportion of difficult words in a text (lexical sophistication; Kyle & Crossley, 2015; Laufer & Nation, 1995; Meara, 2005). What connects the aforementioned constructs is the goal for which they are generally used, which is (broadly speaking) to measure productive lexical proficiency.

This chapter represents a critical introduction to a number of methods for measuring lexical richness (density, diversity, and sophistication). Key theoretical and practical issues are introduced, and a sample text produced by an L2 writer is used to demonstrate how each index of lexical richness is calculated. Additionally, a number of freely available text analysis tools for the measurement of lexical richness are introduced.

Critical Issues and Topics

Preliminary Vocabulary

Before discussing the various aspects of lexical richness, it is important to define some key terms related to how lexical items can be defined, as the chosen definitions can seriously impact the results of a study.

Types and Tokens

The number of *tokens* in a text refers to the number of running words in a text. For example, the sample text in Table 29.1 comprises 73 tokens. The number of *types* in a text refers to

Table 29.1 Sample text in three formats: original, tokenized, and typed

Original text	How do you feel if you enter a restaurant full of smoke? I want to escape from it as soon as possible. So in my opinion, smoking should be completely banned at all the restaurants in the country. As we all know, restaurant is a public place where people go there for enjoying meals. If someone there smokes, the environment is polluted. Can you enjoy a meal at a place full of smoke?
Tokens = 73	how do you feel if you enter a restaurant full of smoke i want to escape from it as soon as possible so in my opinion smoking should be completely banned at all the restaurants in the country as we all know, restaurant is a public place where people go there for enjoying meals if someone there smokes, the environment is polluted can you enjoy a meal at a place full of smoke
Word types = 54 type (frequency)	a (4), as (3), the (3), you (3), all (2), at (2), full (2), if (2), in (2), is (2), of (2), place (2), restaurant (2), smoke (2), there (2), banned (1), be (1), can (1), completely (1), country (1), do (1), enjoy (1), enjoying (1), enter (1), environment (1), escape (1), feel (1), for (1), from (1), go (1), how (1), i (1), it (1), know (1), meal (1), meals (1), my (1), opinion (1), people (1), polluted (1), possible (1), public (1), restaurants (1), should (1), smokes (1), smoking (1), so (1), someone (1), soon (1), to (1), want (1), we (1), where (1)

Note: Text from the International Corpus Network of Asian Learners of English (Ishikawa, 2011).

the number of *unique* words in a text. The sample text in Table 29.1 includes 54 word types. It should be noted that the operational definition of a lexical item will affect type counts in most cases (see Table 29.1).

Words, Lemmas, Flemmas, and Families

When calculating indices of lexical richness, it is important to consider the operational definition of a lexical item. Such a definition should reflect the presumed lexical knowledge of the writers/speakers. In most cases, lexical items are preliminarily identified as strings of letters separated by white space, end punctuation, hyphens, or commas. From there, lexical items can be grouped (or not) in a number of ways. The most common operational definitions of lexical items are included later in the chapter. The consequences of these choices will be illustrated with regard to type counts and reference corpus frequency counts in reference to the sample sentences in Table 29.2.

When lexical items are operationally defined as **words**, each distinct word form is treated as a separate unit. For example, each of the inflected and derived forms of the word *smoke* (e.g., *smoke*, *smoker*, *smoking*, etc.) would be treated as a distinct word. Defining lexical items as words creates an implicit presumption that writers/speakers have little to no knowledge of derivational or inflectional morphology. While the former may be true of a wide range of L2 users, the latter is likely only true for very low proficiency L2 users (i.e., true beginners).

A **lemma** is a group of words that share both semantics and part of speech. Thus, all inflected forms of a word are treated as members of the same lemma form (e.g., all inflections of the verb *to smoke*, including *smoke*, *smokes*, *smoked*, and *smoking* are members of the same lemma). In Table 29.2 we see that the type counts for lemmas differ slightly from the word type counts. In the lemmatized version of the text, both *be* and *is* are counted as the same form (the lemma "be"). Similarly, the singular *restaurant* and plural *restaurants* are

Table 29.2 Words, lemmas, and families in the example text

Word types = 53 type (frequency)	a (4), as (3), the (3), you (3), all (2), at (2), full (2), if (2), in (2), is (2), of (2), place (2), restaurant (2), smoke (2), there (2), banned (1), be (1), can (1), completely (1), country (1), do (1), enjoy (1), enjoying (1), enter (1), environment (1), escape (1), feel (1), for (1), from (1), go (1), how (1), i (1), it (1), know (1), meal (1), meals (1), my (1), opinion (1), people (1), polluted (1), possible (1), public (1), restaurants (1), should (1), smokes (1), smoking (1), so (1), someone (1), soon (1), to (1), want (1), we (1), where (1)
Lemma types = 48 type (frequency)	a (4), as (3), be (3), restaurant (3), the (3), you (3), all (2), at (2), enjoy (2), full (2), i (2), if (2), in (2), meal (2), of (2), place (2), smoke_noun (2), there (2), ban (1), can (1), completely (1), country (1), do (1), enter (1), environment (1), escape (1), feel (1), for (1), from (1), go (1), how (1), it (1), know (1), opinion (1), people (1), polluted (1), possible (1), public (1), should (1), smoke_verb (1), smoking_noun (1), so (1), someone (1), soon (1), to (1), want (1), we (1), where (1)
Family types = 46 type (frequency)	a (4), smoke (4), as (3), be (3), restaurant (3), the (3), you (3), all (2), at (2), enjoy (2), full (2), i (2), if (2), in (2), meal (2), of (2), place (2), there (2), ban (1), can (1), complete (1), country (1), do (1), enter (1), environment (1), escape (1), feel (1), for (1), from (1), go (1), how (1), it (1), know (1), opinion (1), people (1), pollute (1), possible (1), public (1), should (1), so (1), some (1), soon (1), to (1), want (1), we (1), where (1)

Note: Text from ICNALE corpus, W_CHN_SMK_022_A2_0.txt; families defined as per the revised BNC_COCA lists available on www.laurenceanthony.net.

counted as the same form (the lemma *restaurant*). Although these frequency differences are relatively few in our small sample, they can have a much larger effect over an entire corpus (particularly for verb frequencies). The use of lemmas indicates an implicit presumption that learners have a working knowledge of the verb and noun inflection system. As noted earlier, this is likely true for all but the lowest proficiency L2 users.

A slight variation of the lemma is the **flemma** (Nation, 2016; Pinchbeck, 2014, 2017), which includes all inflected forms of a words and their homographs (i.e., while lemmas are sensitive to part of speech, flemmas are not). In a flemmatized version of the sample text, the noun *smoke* and the verb *smoke* would both be represented as the same flemma *smoke*. Based on a strict definition, the noun *smoking* would also be counted as part of the flemma *smoke*. A practical advantage to the use of flemmas is that flemmatization of learner texts can be done without the use of a part of speech tagger, and in fact many tools that purport to lemmatize texts actually flemmatize them.

A word **family** (Nation, 2001) includes all inflected forms of a word and most of its derived forms. Most studies include derived forms up to and including Level 6 on Bauer and Nation's (1993) word family hierarchy. In Table 29.2, we see that the three lemmas related to the base form of *smoke* (the noun *smoke*, the noun *smoking*, and the verb *smoke*) are all counted as a single family, raising the frequency count to four instances. The use of families indicates an implicit presumption that learners have working knowledge of the inflection and derivation systems of a language. Although this may be a reasonable presumption for advanced L2 users and particularly in receptive modes (Nation & Webb, 2011; Webb, 2010) it may not be appropriate for lower proficiency L2 users and for productive modes (McLean, 2018; Schmitt, 2010).

Function and Content Words

Distinction between content words and function words is made in the calculation of a number of indices related to lexical richness. *Function* words refer to closed-class, grammatical words such as articles and prepositions (among many others). *Content* words refer to open-class, lexical words, including nouns, adjectives, most verbs, and some adverbs (Engber, 1995; Lu, 2012; Quirk, Greenbaum, Leech, & Svartik, 1985). Lexical verbs tend to be counted as content words, while copular verbs (i.e., *to be*) and auxiliary verbs (e.g., modals and auxiliary uses of *to be*, *to have*, and *to do*) tend to be counted as function words. Content adverbs tend to include those with adjectival bases (e.g., adverbs that end in *–ly* such as *basically* and *slowly*) while all other adverbs tend to be counted as function words (c.f., O'Loughlin, 1995). Table 29.3 includes a list of content and function lemma types that occur in the sample sentences.

Measuring Lexical Density

Of the three aspects of lexical richness that are outlined in the literature, lexical density is likely the one that is employed least often in productive L2 vocabulary research. Indices of lexical density compare the number of content words to the number of total words in a text. In doing so, lexical density provides a measure of the density of information in a text (Ure, 1971; Halliday, 1985). As we can see from Table 29.1 and Table 29.3, our sample text has 73 tokens, of which 31 are content words. Lexical density for our text is therefore calculated as:

$$lexical\ density = \frac{number\ of\ content\ words}{number\ of\ total\ words} = \frac{31}{73} = 0.425$$

Research has indicated that lexical density is a useful metric for distinguishing between spoken and written modes (Ure, 1971) and more specifically between performances on tasks that are more or less interactive (O'Loughlin, 1995; Shohamy, 1994). These studies indicated that less interactive tasks (including writing tasks) tend to have higher lexical density than more interactive ones (such as interviews). A number of studies have also attempted to use lexical density to model writing and speaking proficiency (Engber, 1995; Hyltenstam, 1988; Linnarud, 1986; Lu, 2012; Nihalani, 1981). Although two studies (Linnarud, 1986; Engber, 1995) demonstrated a small effect between lexical density and writing proficiency, no study that I am aware of has reported a significant relationship between the two. This

Table 29.3 Content and function lemmas that occur in the sample sentences

Content lemmas 31 tokens 24 types	restaurant (3), enjoy (2), full (2), meal (2), place (2), smoke_noun (2), ban (1), completely (1), country (1), enter (1), environment (1), escape (1), feel (1), go (1), know (1), opinion (1), people (1), polluted (1), possible (1), public (1), smoke_verb (1), smoking_noun (1), someone (1), want (1)
Function lemmas 42 tokens 24 types	a (4), as (3), be (3), the (3), you (3), all (2), at (2), i (2), if (2), in (2), of (2), there (2), can (1), do (1), for (1), from (1), how (1), it (1), should (1), so (1), soon (1), to (1), we (1), where (1)

preliminarily suggests that while lexical density may be appropriate for some purposes (e.g., distinguishing modes or registers), it may not be appropriate for measuring productive lexical proficiency.

Measuring Lexical Diversity

Lexical diversity indices are used to measure the variety of lexical items in a text, which is presumed to reflect the extent of the lexical knowledge of the writer of that text. Because more proficient speakers and writers have more lexical items at their disposal, they are able to use a wide variety of lexical items to accomplish a task, while less proficient speakers and writers will tend to repeat a smaller number of lexical items (e.g., Read, 2000). The most basic way to calculate lexical diversity is the simple type-token ratio (TTR; i.e., the number of types in a text divided by the number of tokens in a text). While some measures such as TTR are relatively straightforward, there are at least three key issues that must be addressed when measuring lexical diversity. First, a researcher must decide whether to analyze lemmas, words, or some other unit (e.g., word families, multiword units, etc.). Second, it must be decided whether all lexical items will be considered, only content lexical items, or only particular parts of speech (e.g., verbs or nouns). The third issue relates to the choice of a diversity index. TTR is consistently and strongly correlated with text length, wherein longer text samples tend to include less diversity (e.g., Linnarud, 1986; McCarthy & Jarvis, 2007). A number of lexical diversity measures have been proposed that attempt to disambiguate text length and diversity scores using statistical transformations and other, more complex methods. However, subsequent research has indicated that many of these, such as Root TTR (also referred to as Guiraud's index; Guiraud, 1960), are strongly correlated with text length (McCarthy & Jarvis, 2007). Next, the simple iteration of TTR is described for reference, followed by a number of indices that have been shown to be relatively independent of text length.

Simple Type-Token Ratio

The simple type-token ratio (TTR) was first proposed by Johnson (1944), and has been used in a number of studies to measure the diversity of words in a text (e.g., Laufer, 1991). TTR is calculated as the number of unique words (types) in a text divided by the number of running words (tokens) in a text: $TTR = \frac{ntypes}{ntokens}$. TTR can be calculated with raw words or lemmas, and can also be confined to all words, content words, or particular parts of speech such as verbs (Harley & King, 1989) or nouns (McClure, 1991). In our sample text (see Table 29.1), there are 48 lemma types and 73 tokens, resulting in $TTR = \frac{48}{73} = .658$. Although I have not found any studies that have done so, it is also possible to calculate TTR with word families. Although TTR is a well-known measure that has been used in some studies (e.g., Laufer, 1991), it is often avoided due to its well-documented negative relationship with text length (e.g., Johnson, 1944; McCarthy & Jarvis, 2007; 2010). Some studies have controlled for length by using a fixed number of words (e.g., the first 50 words) from a text (Grant & Ginther, 2000; Jarvis, Grant, Bikowski, & Ferris, 2003). However, alternatives such as Root TTR, MTLD, and D (see below) are much more commonly employed in the literature.

Maas' Index

Maas' index (Maas, 1972) is a transformation of type-token ratios that attempts to fit the TTR value to a logarithmic curve. This is calculated as: $Maas'\,index = \dfrac{Log(ntokens) - Log(ntypes)}{Log(ntokens)^2}$.

In the sample text, $Maas'\,index = \dfrac{Log(73) - Log(48)}{Log(73)^2} = \dfrac{1.863 - 1.681}{3.472} = .052$. As the ratio of types to tokens increases, Maas' index values decrease. McCarthy & Jarvis (2007, 2010) indicate that Maas' index is relatively independent from text length for spoken genres (r = .320) and particularly for written genres (r = .120). I am currently not aware of any study that has used Maas' index to investigate the relationship between lexical diversity and L2 proficiency, and this may be an area for future investigation.

Moving Average Type-Token Ratio

Moving average type-token ratio (MATTR) is a variant of simple TTR that controls for text length (Covington & McFall, 2010) by taking the average TTR value for all overlapping segments of the text of a specified length (e.g., 50 words). For example, in a 300-word text, the TTR value is calculated for words 1 through 50 in the text, then for words 2 through 51, then for words 3 through 52, and so on, until all of the words have been considered. While more difficult to calculate by hand than simple TTR or Guiraud's index, MATTR is still conceptually straightforward (and easy to interpret), is independent of text length (Covington & McFall, 2010), and is easy to automate. To calculate MATTR in the sample text, we first calculate TTR for the first 50-word segment, which begins with the word *how* and ends with the word *go* (TTR = .780). We then calculate TTR for the next 50-word segment, which begins with the word *do* and ends with the word *there* (TTR = .780), and then for all remaining segments. In total, the sample text includes 24 overlapping 50-word segments that have a mean TTR value of MATTR = .769.

A related variant is the mean segment type-token ratio (MSTTR) wherein the text is divided into non-overlapping segments of n words (e.g., 50 words) and TTR is calculated for each chunk and then averaged across chunks (Engber, 1995; Johnson, 1944). For example, a 250-word text would be divided into five chunks (words 1–50, 51–100, 101–150, 151–200, and 201–250). One problem with MSTTR is that a set of texts are rarely divisible by the same segment length, leaving wasted data. For example, in our example text, only the first 50 words would be used, and the final 23 words would be ignored.

Measure of Textual Lexical Diversity

The measure of textual lexical diversity (MTLD) measures lexical diversity by calculating the average number of words in a text it takes to reach a predetermined TTR value at which TTR is assumed to be stable. McCarthy and Jarvis (2010) analyzed a large corpus of written narrative and expository texts and found that TTR values tended to stabilize at a value of .720. McCarthy and Jarvis (2010) calculated MTLD using successive, non-overlapping text segments of at least ten words (which are referred to as factors). Any leftover segments (i.e., the end of a text) were counted as partial factors. Texts with higher MTLD values can be considered more lexically diverse (because it takes more words on average for TTR values to fall below .720). In the sample text, for example, it takes 67 words for the TTR value to fall below .720. MTLD can also be measured using a moving windows approach and partial factors can be avoided by using the beginning of the text to finish calculating the final factor.

Research has indicated that MTLD is relatively independent of text length (McCarthy & Jarvis, 2010), even for texts as short as 100 words in length (Koizumi & In'nami, 2012). Research investigating the relationship between language proficiency and MTLD scores has indicated a positive relationship between the two. Crossley, Salsbury, and McNamara (2009), for example, found that L2 speakers produced oral samples with higher MTLD values during one-year of study in an intensive English program. A number of other studies have also used MTLD to measure writing quality and lexical proficiency (e.g., Crossley, Salsbury, McNamara, & Jarvis, 2011; Guo, Crossley, & McNamara, 2013). However, these studies also used the index D (see next section), which was more strongly associated with quality/proficiency, and the results for MTLD were not reported.

D

The diversity index D (Malvern et al., 2004; Malvern & Richards, 1997; McCarthy & Jarvis, 2007) is an index with an interesting history. Malvern et al. (2004) conceptualized D as a curve fitting index wherein random samples from 35 to 50 words in length are taken from a text. TTR values for each segment are then used to create an empirical curve, which is then fit to a theoretical curve using a coefficient referred to as D (this is referred to as vocD; see Malvern et al., 2004, for more information). McCarthy and Jarvis (2007), however, found that vocD was essentially an approximation of the sum of probabilities that the words occurring in text would be included in a random sample from the text (in their data, the two indices were correlated at $r = .971$). They proposed the index HD-D, which uses the hypergeometric distribution to determine the probability of a particular word in a text occurring at least once in a random sample of a particular length (in this case, the length is 42 words – an approximate mean value between Malvern et al.'s 35 and 50). HD-D represents the sum of the probabilities that each type in a text will occur in a 42-word sample at least once. Texts with lower HD-D values can be considered to be more diverse than those with higher values. HD-D values can also be converted to the same scale as TTR by multiplying the probability of occurrence of each type by 1/sample size (McCarthy and Jarvis refer to this conversion as average TTR or ATTR). In the sample text, for example, the probability that the word *a*, which occurs four times in the sample will occur at least once in a random sample of 42 words (of our 73-word text) is .971 (i.e., there is a 97.1% chance that *a* would occur at least once). See Table 29.4 for a summary of the HD-D calculation for the sample text.

Table 29.4 HD-D calculation summary for sample text

Word frequency	*Probability word will occur in 42-word sample*	*Words in sample text with frequency*	*Sum of probabilities*
4	.971	1	.023
3	.923	5	.110
2	.823	12	.235
1	.575	30	.411
			HD-D = 32.712
			ATTR = .789

McCarthy and Jarvis (2007) found that vocD (and by extrapolation, HD-D) was relatively independent of text length in written (r = .190) and spoken (r = .350) genres. No studies that I am aware of have used HD-D to measure productive L2 proficiency, likely because until recently no freely available text processing software calculated the measure. VocD, which has been implemented in multiple tools, however, has been used in a number of studies to measure L2 productive proficiency. Guo et al. (2013), for example, found a positive relationship (r = .415) between vocD and holistic TOEFL independent writing scores. Crossley, Salsbury, and McNamara (2012) also found a positive relationship (r = .583) between vocD and written lexical proficiency scores.

Other Approaches

Another issue with prototypical indices of lexical diversity is that they treat all words equally. The sentence *This is my car and I like it*, for example, earns the same diversity score (TTR = 1.0) as the sentence *This exquisite vehicle is a veritable engineering marvel*, despite the fact that the former sentence arguably represents a higher level of lexical proficiency. At least two methods can be used to address this issue. One approach, proposed by Daller, Van Hout, and Treffers-Daller (2003) is to calculate lexical diversity indices that use the number of "advanced" types in the text instead of the total type count. Daller et al. (2003) propose Advanced TTR $\left(\frac{\textit{number of ``advanced'' types}}{\textit{total number of tokens}}\right)$ and Advanced Guiraud $\left(\frac{\textit{number of ``advanced'' types}}{\sqrt{\textit{total number of tokens}}}\right)$. As with TTR and Root TTR, however, these indices will be strongly correlated with text length (Kojima & Yamashita, 2014) and are thus not particularly attractive options. A second method is to use a multivariate approach to the measurement of lexical richness by using both indices of diversity and indices of sophistication (Jarvis, 2013) in multivariate statistical models (e.g., multiple regression). The suitability of such an approach depends on the theoretical and empirical characteristics of the indices used in such a model.

Measuring Lexical Sophistication

Lexical sophistication is often referred to quite specifically as the "number of low frequency words that are appropriate to the topic and style of writing" (Read, 2000, p. 200). This definition highlights the importance that reference-corpus frequency has played in defining the construct. Frequency has been shown to be a reliable predictor of lexical proficiency in a number of studies (Guo et al., 2013; Laufer & Nation, 1995; Kyle & Crossley, 2015). However, a broader definition of lexical sophistication refers to the relative difficulty of learning and/or using a lexical item (e.g., Bulté & Housen, 2012; Kyle, Crossley, & Berger, 2018). This broader definition includes frequency as an important factor in learning (e.g., what makes a word "difficult" or "advanced"), but also recognizes that features beyond frequency (e.g., saliency and contextual distinctiveness among others) affect the learnability of a lexical item (see Ellis, 2002). Research that either implicitly or explicitly follows this latter definition has indicated that the most robust results are obtained using a multivariate approach (e.g., Crossley et al., 2011; Kyle & Crossley, 2015). A number of factors that affect perceptions of sophistication are outlined later in the chapter.

Frequency

Frequency is likely the most widely used method of measuring lexical sophistication. Generally speaking, frequency is measured by comparing the lexical items in a written or spoken text with their frequency in a reference corpus. Lexical items that occur less frequently in the reference corpus tend to be considered more sophisticated (Laufer & Nation, 1995; Guo et al., 2013; Kyle & Crossley, 2015). Spoken and written texts that include higher proportions of less frequent words are considered to be more sophisticated than those that contain higher proportions of more frequent words. As with indices of lexical diversity, decisions must be made with regard to the definition of a lexical item (words, lemmas, families, etc.) and the types of words considered (all words, content words, function words). Additionally, the reference corpus that is used to derive frequency counts is particularly important. A mismatch between reference corpus and target text type (e.g., using a corpus of sitcom subtitles to measure sophistication in academic texts) may result in misleading outcomes. A number of methods have been derived to measure frequency in texts. The most common of these are described next.

Lexical Frequency Profile

The lexical frequency profile (LFP; Laufer, 1994; Laufer & Nation, 1995) method of calculating frequency was one of the earliest approaches to measuring frequency in learner texts. The LFP method divides words in a reference corpus into frequency bands and then calculates the percentage of words in the target text that occur in each band. The LFP method considers word families to be a lexical unit, and all words in a text are analyzed. The original LFP was based on West's (1953) General Service List, though more recent versions include lists based on the British National Corpus (BNC Consortium, 2007) and the Corpus of Contemporary American English (COCA; Davies, 2010). Originally, LFPs distinguished between words that were among the 1,000 most frequent (1K) words, the second most frequent 1,000 words (2K), words that occurred in an early academic word list (Xue & Nation, 1984), and words that occurred in none of the preceding lists (off-list words). After the development of the Academic Word List (AWL; Coxhead, 2000), the academic word portion of LFPs were calculated using the updated list. More recent versions also allow for more fine-grained analyses (e.g., 100-word bands; Cobb, 2017). Table 29.5 shows the percentage of the sample text that are in the 1K, 2K, and AWL lists.

Table 29.5 Lexical frequency profile for the sample text

List	*Percent coverage*	*Cumulative coverage*	*Family headwords*
1K	83.56	83.56	a (4), as (3), be (3), the (3), you (3), all (2), at (2), enjoy (2), full (2), i (2), if (2), in (2), of (2), place (2), there (2), can (1), complete (1), country (1), do (1), enter (1), escape (1), feel (1), for (1), from (1), go (1), how (1), it (1), know (1), opinion (1), people (1), possible (1), public (1), should (1), so (1), some (1), soon (1), to (1), want (1), we (1), where (1)
2K	12.33	95.89	meal (2), restaurant (3), smoke (4)
AWL	1.37	97.26	environment (1)
Off list	2.74	100.00	ban (1), pollute (1)

LFPs have been used in a number of studies to measure lexical sophistication in student writing (Crossley, Cobb, & McNamara, 2013; Laufer, 1994; Laufer & Nation, 1995). These studies have indicated that more proficient L2 users tend to use fewer 1K words than less proficient L2 users. Although the results of an LFP analysis are easy to interpret (and nice visualization of the results is available when using VocabProfile), they have some limitations.

First, LFPs result in a number of related scores instead of a single score, which can be statistically unwieldy. The researcher has to choose which band (or collection of bands; Laufer, 1995) is appropriate for distinguishing between the proficiency levels represented in a particular sample. Second, LFPs are large-grained, meaning that the difference between the 1,000th most frequent word and the 1,001st most frequent word is considered the same as the difference between the most frequent word and the 2,000th most frequent word. This suggests that small gains in vocabulary use are likely to be obscured. Despite these potential limitations, LFPs have been used in a number of studies to model productive lexical proficiency.

P_Lex and S

P-Lex (Meara & Bell, 2001) and S (Kojima & Yamashita, 2014) are broadly based on LFPs, but result in a single score instead of a series of scores making the statistical analyses between frequency and speaking or writing proficiency (broadly construed) much more straightforward. Because both P-Lex and S are based on LFPs, word families are used as the lexical unit, and both content items and function items in a text are included in the analysis. P-Lex is calculated by first dividing a text into 10-item segments. Then, the number of advanced words in each segment (defined as words that are not among the 1K list, are not proper nouns, and are not numbers) is counted. The proportion of segments that include no advanced words, one advanced word, two advanced words and so on is calculated. The resulting data is then fit to a Poisson distribution (i.e., a distribution that is skewed to the left) using lambda. Texts with a low lambda score will have a steeper curve (i.e., will include fewer advanced words) and therefore be considered less sophisticated than those with a higher lambda score. To calculate lambda, our sample text is divided into seven segments. Five of these segments include one advanced word, while two of the segments include two advanced words. Our sample text earns a lambda score of 1.285 (see Figure 29.1 for a visualization of these results).

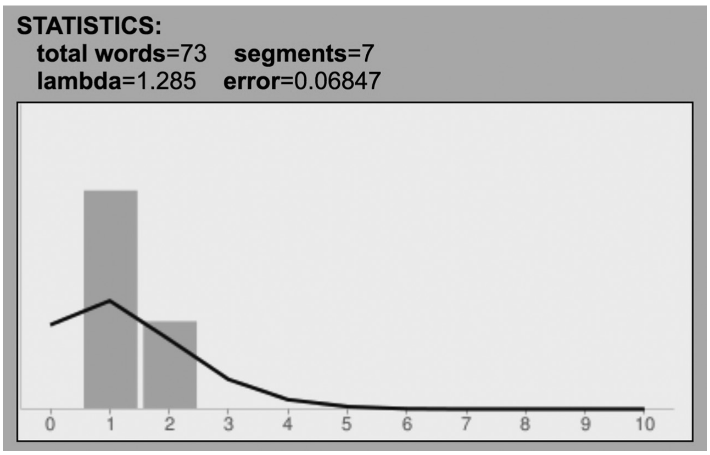

Figure 29.1 Text segments containing ~~advanced x number of~~ advanced words

One potential weakness of P_Lex is the binary distinction that is made between basic words and advanced words, and thus a word that is the 1,001st most frequent word is considered to be as advanced as the 5,000th most frequent word. This may work well for distinguishing between low- and high-proficiency learners, but may not be appropriate for distinguishing between higher proficiency learners.

The S procedure, which was proposed by Kojima and Yamashita (2014), potentially improves upon P_Lex by considering finer-grained frequency bands. The index S is calculated by determining the cumulative text coverage of six 500-word frequency bands (1–500, 501–1,000, 1,001–1,500, 1,501–2,000, 2,001–2,500, and 2,501–3,000) in successive, overlapping 50-word segments in a text. The coverage rate is then fit to a theoretical curve, and S represents the point at which cumulative coverage would hit 100%. An S-value of 2,555, for example, would suggest that all words used in a text would be covered using the 2,555 most frequent word families. Higher S values suggest that, on average, writers and speakers are using less-frequent words. According to the S calculator on Kojima's website, the sample text earns an S score of 2978.42, suggesting that the writer of our sample text has a productive vocabulary of approximately 3,000 words. One limitation of S is that there is currently no available tool that can calculate the measure for more than one text at a time. If such a tool becomes available, S may be an attractive option for the calculation of frequency scores.

Mean Frequency

Another method of calculating frequency scores is to calculate the average frequency score for either all lexical items, content lexical items, or function lexical items in a text. Two advantages to using mean frequency scores as opposed to band scores is that they provide a single score (which eases further statistical analyses) and they are fine-grained (i.e., precise differences in word frequencies are maintained). Often, content frequency is differentiated from function frequency because function items tend to be of particularly high frequency. Some variants of mean frequency scores include logarithmically transformed frequency scores and mean rank frequency scores. Table 29.6 includes mean frequency scores for content flemmas and function flemmas in our sample text.

Mean frequency scores (particularly for content words) have demonstrated a fairly consistent negative relationship with writing and speaking proficiency scores, wherein performances perceived to be of higher quality tend to have lower mean frequency values (e.g., Guo et al., 2013; Kyle & Crossley, 2015; Kyle et al., 2018). A

Measuring Lexical Sophistication: Beyond Frequency

While frequency has been a mainstay in the measurement of lexical sophistication for some time (e.g., Laufer, 1994; Laufer & Nation, 1995), related research in psychology (e.g., Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Paivio, Yuille, & Madigan, 1968) and language development (Crossley, Kyle, & Salsbury, 2016; Ellis, 2006; MacWhinney, 2008) has suggested that frequency does not fully account for what makes a word more difficult to learn or use. Other lexical features such as saliency (e.g., concreteness), entrenchment (e.g., lexical access), and contextual distinctiveness (e.g., range), among many others, can be used in concert with frequency to more accurately model the sophistication of a particular lexical item. Further, a number of studies (Crossley et al., 2012; Guo et al., 2013; Kim, Crossley, & Kyle, 2018; Kyle & Crossley, 2015) have indicated the importance of measuring lexical sophistication using multivariate methods and as

Table 29.6 Frequencies per million words in the sample text based on the magazine section of COCA

Content flemma	*Frequency*	*Function flemma*	*Frequency*
go (1)	1,820.769	the (3)	53,641.466
people (1)	1,403.408	be (3)	32,439.882
know (1)	1,298.598	a (4)	28,259.271
want (1)	1,004.511	of (2)	25,782.247
place (2)	710.293	to (1)	24,925.015
feel (1)	669.088	in (2)	18,141.039
country (1)	454.233	for (1)	9,082.281
public (1)	349.851	you (3)	8,826.562
full (2)	222.047	i (2)	8,445.369
possible (1)	211.787	it (1)	8,260.920
someone (1)	204.881	as (3)	6,145.724
enjoy (2)	159.310	we (1)	5,480.012
enter (1)	129.806	at (2)	4,945.371
environment (1)	112.475	do (1)	4,751.696
restaurant (3)	103.513	from (1)	4,282.254
meal (2)	97.882	can (1)	3,012.655
completely (1)	94.473	if (2)	2,361.667
escape (1)	62.121	all (2)	2,299.371
opinion (1)	57.249	so (1)	2,043.916
smoke (3)	53.917	there (2)	1,960.505
ban (1)	40.963	how (1)	1,226.569
smoking (1)	27.778	where (1)	936.540
polluted (1)	4.553	should (1)	714.043
		soon (1)	196.677
Mean frequency (types)	404.066		10,756.710
Mean frequency (tokens)	469.474		25,374.237

Note: Counts are based on the downloadable version of COCA that includes texts from 1990 to 2012. Because the frequency counts are based on flemmas, the reported frequency of *smoke* includes counts for nouns, adjectives, and verbs.

a multidimensional construct. A non-exhaustive list of indices of lexical sophistication are included later, including word recognition norms, psycholinguistic word information, and contextual distinctiveness. Further description of indices and descriptions of a wider range of indices can be found in Kyle et al. (2018).

Psycholinguistic Word Information

Psycholinguistic word information indices measure word characteristics beyond frequency. These norms refer to individuals' perceptions of each characteristic for a particular word based on behavioral studies (i.e., surveys). One commonly used index is concreteness, which refers to how abstract the concept denoted by a word is and may be linked to the concept of saliency (Crossley et al., 2016). More concrete words are theorized to be easier to learn (Paivio, 1971; Schwanenflugel, Harnishfeger, & Stowe, 1988), and accordingly texts that on average include less concrete words tend to earn higher proficiency scores (e.g., Kyle et al., 2018). Other related psycholinguistic word information indices include familiarity, imageability, meaningfulness, and age of acquisition (Coltheart, 1981; Kyle et al., 2018; McNamara, Crossley, & McCarthy, 2010). Table 29.7 includes concreteness scores collected by Brysbaert et al. (2014)

Table 29.7 Average concreteness score for words in the sample text

Content flemma	*Concreteness*	*Function flemma*	*Concreteness*
smoke (3)	4.96	the (3)	1.43
restaurant (3)	4.89	be (3)	1.85
people (1)	4.82	a (4)	1.46
meal (2)	4.62	of (2)	1.67
smoking (1)	4.32	to (1)	1.55
country (1)	4.17	in (2)	3
environment (1)	3.74	for (1)	1.63
someone (1)	3.71	you (3)	4.11
full (2)	3.59	i (2)	3.93
place (2)	3.48	it (1)	2.81
escape (1)	3.34	as (3)	1.33
polluted (1)	3.34	we (1)	3.08
go (1)	3.15	at (2)	2.07
enter (1)	3.12	do (1)	2.46
public (1)	2.57	from (1)	1.84
ban (1)	2.37	can (1)	4.55
enjoy (2)	2.29	if (2)	1.19
feel (1)	2.28	all (2)	2.27
completely (1)	2.18	so (1)	1.42
want (1)	1.93	there (2)	2.2
opinion (1)	1.93	how (1)	1.35
know (1)	1.68	where (1)	1.66
possible (1)	1.56	should (1)	1.53
		soon (1)	1.79
Mean concreteness score (types)	3.219		2.174
Mean concreteness score (tokens)	4.683		3.764

Source: Based onBrysbaert et al., 2014

for each flemma in the sample text. Flemmas such as *smoke*, *restaurant*, and *people* are highly concrete, while flemmas such as *possible* and *know* are much less concrete.

Word Recognition Norms

Word recognition norms, such as lexical decision latencies, measure how quickly an individual can access a word. In a lexical decision task, individuals are presented with a string of characters (e.g., letters) and must decide whether that string is a word or not in the target language (e.g., English). Words that can be identified as part of the target language more quickly are suggested to be more entrenched in an individual's memory, and would therefore be considered less sophisticated than words that can be identified less quickly. Table 29.8 includes average lexical decision times (measured in milliseconds) for the words in the sample text based on data collected by Balota et al. (2007). Words such as *people*, *escape*, and *know* are accessed quickly (in approximately .5 seconds), while words such as *banned*, *polluted*, and *environment* are accessed more slowly (up to .8 second for the latter). Although

Table 29.8 Average lexical decision time (in milliseconds) for words in the sample text

Content word	*Lexical decision time*	*Function word*	*Lexical decision time*
people (1)	540.85	to (1)	551.27
escape (1)	546.38	so (1)	555.56
know (1)	549.71	there (2)	556.39
meal (1)	562.89	do (1)	558
restaurants (1)	568.97	we (1)	561.34
enter (1)	576.53	it (1)	566.85
enjoy (1)	582.29	soon (1)	585.28
place (2)	588.84	is (2)	588.85
feel (1)	593.65	my (1)	590.16
want (1)	596.21	all (2)	592.33
full (2)	597.63	for (1)	596.19
go (1)	597.69	can (1)	596.41
smoking (1)	607.32	be (1)	598.42
restaurant (2)	608.61	how (1)	599.37
smoke (2)	609	the (3)	600.19
public (1)	611.16	from (1)	607.85
possible (1)	623.06	should (1)	607.91
opinion (1)	623.69	you (3)	615.21
country (1)	633.18	in (2)	617.97
enjoying (1)	639.58	of (2)	620.42
completely (1)	640.97	i (1)	631.28
someone (1)	641.15	where (1)	635.06
smokes (1)	663.64	at (2)	651.8
banned (1)	733.69	as (3)	653.09
meals (1)	733.69	if (2)	658.3
polluted (1)	792.19	a (4)	798.92
environment (1)	816.82		
Mean lexical decision time (types)	625.16		607.48
Mean lexical decision time (tokens)	714.20		1008.24

Note: Balota et al. report lexical decision times for words (not lemmas or families).

word recognition norms have only recently begun to be used in the measurement of lexical sophistication, preliminary results are promising (see, e.g., Berger, Crossley, & Kyle, 2019).

Contextual Distinctiveness

Indices of contextual distinctiveness measure the degree to which a word is used in restricted context or is used in a wide variety of contexts. Words that are used in a wider variety of contexts are more likely to be encountered frequently, and therefore on average will be learned earlier/more easily than more specialized words. Contextual distinctiveness is measured using a number of methods. The most straightforward is range, which is calculated as the number of texts in a corpus a particular word occurs in, and is often reported as a percentage. For example, *the* occurs in 99.85% of texts in the magazine section of COCA while *smoke* occurs in only 5.67% of the texts. Another more statistically advanced corpus method is called

SemD (Hoffman, Ralph, & Rogers, 2013), which determines the number of contexts a word will occur in based on latent semantic analysis (Landauer, Foltz, & Laham, 1998). Words such as *possible* earn higher SemD values than words such as *polluted*, suggesting *possible* will occur in more contexts than *polluted*. A third method measures the number of words a particular lexical item is associated with based on free word association tasks, wherein participants are given a particular lexical item and are asked to list as many words as they can that are associated with that item (e.g., Kiss, Armstrong, Milroy, & Piper, 1973; Nelson, McEvoy, & Schreiber, 2004). Lexical items that elicit more associated words (and words that are elicited more often by other words) are less contextually distinct than those that elicit (or are elicited by) fewer. For example, Nelson et al. (2004) report that words such as *people* were elicited by 154 other words, while *smoking* was only elicited by six other words. Table 29.9 includes

Table 29.9 Number of associated words for each word in the sample text

Use USF			
Content word	*Contextual diversity*	*Function word*	*Contextual diversity*
people (1)	154	do (1)	48
go (1)	70	can (1)	36
country (1)	65	in (2)	21
place (2)	62	to (1)	17
smoke (2)	44	all (2)	15
know (1)	38	you (3)	15
want (1)	33	there (2)	14
opinion (1)	29	be (1)	13
full (2)	28	how (1)	12
restaurant (2)	23	soon (1)	8
meal (1)	21	is (2)	8
feel (1)	21	it (1)	7
environment (1)	13	where (1)	7
escape (1)	12	a (4)	7
enjoy (1)	12	for (1)	6
enter (1)	11	from (1)	6
someone (1)	7	of (2)	6
smoking (1)	6	the (3)	5
public (1)	6	i (1)	4
possible (1)	5	so (1)	3
banned (1)	n/a	my (1)	3
completely (1)	n/a	should (1)	2
enjoying (1)	n/a	if (2)	2
meals (1)	n/a	as (3)	1
polluted (1)	n/a	at (2)	n/a
restaurants (1)	n/a	we (1)	n/a
smokes (1)	n/a		
Mean contextual diversity (types)	33.00		11.08
Mean contextual diversity (tokens)	40.85		16.46

Source: Based on Nelson et al., 2004

elicitation counts for each of the words in the sample text based on Nelson et al. Indices of contextual distinctiveness have been found to be predictive of both written (Kyle & Crossley, 2015) and spoken (Berger, Crossley, & Kyle, 2017) lexical proficiency.

Other Approaches

The indices outlined earlier represent only a small subset of the lexical characteristics that can be measured. Lexical features such as polysemy, hypernymy, and word neighbor indices can all provide further insights into the construct of lexical sophistication (Crossley et al., 2009; Guo et al., 2013; Kyle et al., 2018). One particularly important feature of lexical sophistication that was not discussed earlier and is beyond the scope of this chapter is collocation. As noted by many scholars, one important aspect of productive vocabulary knowledge is knowing which words tend to be used together (Nation, 2001; Nation & Webb, 2011; Sinclair, 1991). Accordingly, a number of studies have demonstrated a strong, positive relationship between the use of associated lexical items and proficient productive lexical use (Bestgen & Granger, 2014; Kyle & Crossley, 2015).

Tools for Automatically Measuring Lexical Richness

One benefit to the measurement of lexical richness is that there are many freely available resources available that calculate various indices of lexical richness. When choosing a tool for the measurement of lexical richness, there are at least four important considerations. The first, most obvious consideration is that the tool calculates the index/indices that you are interested in. Second, with an increasing amount of learner data available (i.e., learner corpora), a convenient feature is batch processing (i.e., the ability to process multiple texts at once). Third, transparency in the calculation of various measures is important. Some tools are black boxes that produce numbers from texts, but no one except the tool developers have the ability to determine whether one's particular texts are being processed properly. A fourth, more minor issue is data security. Sensitive/privileged data should likely not be processed over an unsecure web interface. Six groups of tools are described below with regard to (1) the indices measured, (2) the convenience of using the tool, and (3) transparency.

CLAN

CLAN is a text analysis program developed by Brian MacWhinney to analyze child language development data (MacWhinney & Snow, 1990). CLAN is available on both Windows and Apple operating systems and can be freely downloaded from https://childes.talkbank.org. CLAN calculates a wide range of measures for texts that adhere to CHILDES format (an automatic text converter is available through CLAN), which was designed to record interactions between children and their caretakers. With regard to the measurement of lexical richness, CLAN is likely most useful for its ability to calculate the lexical diversity index vocD. CLAN allows for batch processing and also provides some diagnostic output. CLAN is particularly useful for calculating complexity measures in interaction data, and while it can be used to calculate lexical richness indices, other tools may be easier to use. If for some reason a researcher needs to calculate vocD (as opposed to different instantiations of D such as HD-D), however, CLAN is the best tool for the job.

Coh-Metrix

Coh-metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara, Graesser, McCarthy, & Cai, 2014) has been used in a large number of studies by the research team that developed it. Coh-metrix calculates a wide range of indices related to lexical diversity (e.g., TTR for content words, MTLD, and vocD) and sophistication (e.g., psycholinguistic word information and mean frequency scores). The version of the tool that is available to users outside of the research team (www.cohmetrix.com) only allows for a single text to be processed at a time, which limits its usefulness for all but the smallest datasets. The online version of Coh-metrix also provides no diagnostic output, which also limits its attractiveness.

Lexical Complexity Analyzer

The lexical complexity analyzer (LCA; Lu, 2012) calculates one index of lexical density, 19 indices related to lexical diversity, and five band-based indices related to lexical sophistication. LCA is available in two versions. The first is a freely available Python script that works on Apple and Linux operating systems and is accessed via the command line (this requires a very small amount of computer literacy). With a small amount of programing knowledge, one can manipulate the script to provide any diagnostic information desired. This version can be downloaded from Xiaofe Lu's website (www.personal.psu.edu/xxl13). The second version, which is accessed via an online interface requires no programing knowledge and can batch process up to 200 texts at a time (see Haiyang Ai's website: http://aihaiyang.com/software/lca/). The only potential downside to the lexical complexity analyzer is that it does not calculate any of the lexical diversity indices suggested in this chapter (Maas' index, MATTR, MTLD, or HD-D). However, LCA does include a wide range of indices that have been commonly used in the field of applied linguistics.

P_Lex and S

At the time of writing, P_Lex (Meara & Bell, 2001) is only freely available as an online tool that can be accessed on Paul Meara's webpage (www.lognostics.co.uk). Texts can only be processed in single-file mode, limiting its usefulness for large-scale analyses. It does, however, provide a great deal of information regarding how the index is calculated for each text, which makes it particularly transparent. The related measure S (Kojima & Yamashita, 2014) can also only be calculated in single-file mode on Masumi Kojima's website (www.kojima-vlab.org). Although a number of curve fit diagnostics are reported, no other diagnostic output is provided.

Range, AntWordProfiler, and VocabProfile

Lexical frequency profiles (LFPs) can be generated for word families (with types or tokens) using Range (Heatley & Nation, 1994), which is available on Paul Nation's website, or AntWordProfiler (Anthony, 2014), which is available on Laurence Anthony's website (www.laurenceanthony.net). Range works on Windows machines only, but AntWordProfiler is available for both Windows and Apple operating systems. Both Range and AntWordProfiler allow for batch processing and provide a great deal of diagnostic output. Even more diagnostic output can be provided for single texts using VocabProfile (Cobb, 2018), which can be accessed at www.lextutor.ca/. The only downside to Range and AntWordProfiler is the format of the output. The output for each text is provided as a table (with intervening text between the tables), which requires a fair amount of cleaning before any statistical analyses can be conducted.

TAALED and TAALES

The Tool for the automatic analysis of Lexical Diversity (TAALED) and the related Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015; Kyle et al., 2018) together calculate the widest range of lexical richness indices of any freely available tool. These tools are available on Kristopher Kyle and Scott Crossley's website (www.linguisticanalysistools.org). TAALED calculates lexical density for types and tokens, and eight distinct indices of lexical diversity, including Maas' index, MATTR, MTLD, and HD-D. TAALES calculates over 500 indices related to lexical sophistication including mean word frequency for a variety of reference corpora (for all words, content words, and function words), psycholinguistic word information, word recognition norms, contextual distinctiveness, polysemy, hypernymy, and n-gram association strength (among many others). Both TAALED and TAALES are available for Windows and Apple operating systems and allow for batch processing. Both TAALED and TAALES provide optional detailed diagnostic information for post-hoc analyses such as individual index scores for each word in a text. Additionally, TAALES 3.0, which will be released in late 2019 will also calculate LFPs, lambda (as per P_Lex), and S.

Future Directions

Lexical Diversity

As noted, indices of lexical diversity are widely used to model lexical richness (and, by extrapolation, productive lexical proficiency). However, care must be taken when selecting an index to use, as many commonly used indices are strongly correlated with text length and should be avoided. Preliminarily, the preceding four indices described favorably (Maas' index, MATTR, MTLD, and HD-D) appear to be attractive options in light of their relative independence of text length (e.g., McCarthy & Jarvis, 2007, 2010). At least two issues, however, warrant exploration. First, most studies that have investigated the impact of length (e.g., McCarthy & Jarvis, 2007, 2010) have used text segments from fairly long text (i.e., 2,000 words) to do so. L2 texts that are commonly analyzed in vocabulary studies tend to be much shorter (e.g., between 100 and 500 words), which may affect the relationship between these indices and text length. Koizumi and In'nami (2012), for example, found that only MLTD was relatively stable for texts between 110 and 200 words. However, their sample was fairly small ($n = 38$) and only represented one genre and mode, making this is a particularly rich area for future research. Second, as alluded to earlier, lexical diversity is most often measured as a univariate construct. However, Jarvis (2013) outlines a number of ways in which diversity can be conceptualized as a multivariate construct that includes a number of distinct aspects. A full treatment of Jarvis' ideas is beyond the scope of this chapter, but make an excellent starting point for those interested in further investigating the construct.

Lexical Sophistication

Given the number of available indices, the calculation of lexical sophistication can be quite complicated. For some purposes, the use of straightforward and univariate indices such as a variation of the LFP or mean frequency scores may be appropriate. However, both theoretical perspectives and empirical studies suggest that frequency is not the only determinant in the ease/difficulty of learning and using a word. This suggests that to obtain a fuller understanding of the sophistication of lexical items in a L2 user-produced text, multiple measurements are necessary. There are at least two directions that can be taken to move the field forward with regard to the measurement of lexical sophistication. First, unlike indices of lexical diversity, indices of lexical sophistication have not been subject to widescale, rigorous testing of length dependence, length stability, or reliability (though see Kojima & Yamashita, 2014). This is an important aspect of future research. Second, indices of lexical sophistication need to be tested in a variety of contexts to determine the degree to which these are stable across registers and modes.

Conclusion

This chapter has provided an overview of indices used to measure lexical richness, including those related to lexical density, lexical diversity, and lexical sophistication. The calculation of some of these indices has been outlined, with accompanying examples, and a number of critiques and suggestions have been made. Additionally, a number of freely available tools have been described, and each has benefits and drawbacks that each researcher must weigh. It should be clearly stated that no single index (or set of indices) is the best way to measure lexical richness in every context. For each study, researchers must make informed decisions about the benefits and drawbacks of each index (or set of indices). It is hoped that this chapter has provided the first steps in enabling such an informed choice.

Further Reading

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106.

This article provides a critical examination of the construct of lexical diversity and introduces a multivariate approach for its measurement.

Kyle, K., Crossley, S. A., & Berger, C. M. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046.

This article introduces a wide range of indices related to lexical sophistication. At time of writing, it represents the most up-to-date description of newly developed indices. Importantly, it includes two validation studies that indicate the relationship between these indices and ratings of written lexical proficiency and writing quality.

Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.

This widely cited book provides an excellent introduction to the construct of lexical richness and its measurement.

Related Topics

The relationship between vocabulary knowledge and proficiency, factors affecting the learning of single-word items, factors affecting the learning of multiword items, key issues in measuring vocabulary knowledge, key issues in researching single-word items, key issues in researching multiword items

References

Anthony, L. (2014). *AntWordProfiler* (Version 1.4. 1) [Computer Software]. Tokyo, Japan: Waseda University.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Berger, C. M., Crossley, S. A., & Kyle, K. (2017). Using novel word context measures to predict human ratings of lexical proficiency. *Journal of Educational Technology & Society*, *20*(2), 201–212.

Berger, C. M., Crossley, S. A., & Kyle, K. (2019). Using Native-Speaker Psycholinguistic Norms to Predict Lexical Proficiency and Development in Second Language Production. *Applied Linguistics*, *40*(1), 22–42.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/j.jslw.2014.09.004

BNC Consortium. (2007). *The British National Corpus, version 3*. BNC Consortium. Retrieved from www.natcorp.ox.ac.uk/

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & F. Vedder (Eds.), *Dimensions of L2 performance and proficiency: complexity, accuracy, and fluency in SLA*. Amsterdam: John Benjamins.

Cobb, T. (2017, March). *Broad v. narrow band lexical frequency profiling*. Presented at the American Association for Applied Linguistics, Portland. Retrieved from www.lextutor.ca/vp/comp/vocab@aaal.pdf

Cobb, T. (2018). *Web VocabProfile*. http://www.lextutor.ca/vp/

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type – token ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94–100.

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

Crossley, S. A., Cobb, T., & McNamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Implications for active vocabulary research and pedagogical applications. *System*, *41*(4), 965–981. https://doi.org/10.1016/j.system.2013.08.002

Crossley, S. A., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal*, *100*(3), 702–715.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. https://doi.org/10.1177/0265532211419331

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. https://doi.org/10.1177/0265532210378031

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

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, *25*(4), 447–464. https://doi.org/10.1093/llc/fqq018

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (2006). Selective attention and transfer phenomena in L2 acquisition: Contingency, cue competition, salience, interference, overshadowing, blocking, and perceptual learning. *Applied Linguistics*, *27*(2), 164–194.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155. https://doi.org/10.1016/1060-3743(95)90004-7

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*(2), 123–145.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris: Presses universitaires de France.

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing*, *18*(3), 218–238.

Halliday, M. A. (1985). *Spoken and written language*. Melbourne: Deakin University Press.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition*, *11*(4), 415–439.

Heatley, A., & Nation, I. S. P. (1994). *Range*. Victoria University of Wellington, NZ. [Computer Program]. Retrieved from http://www.vuw.ac.nz/lals/

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods*, *45*(3), 718–730.

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual & Multicultural Development*, *9*(1–2), 67–84.

Ishikawa, S. (2011). A new horizon in learner corpus studies: The aim of the ICNALE project. *Corpora and Language Technologies in Teaching, Learning and Research*, 3–11.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, *63*(s1), 87–106.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377–403. https://doi.org/10.1016/j.jslw.2003.09.001

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, *56*, 1–15.

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods 50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal 102*(1), 120–141.

Kiss, G. R., Armstrong, C., Milroy, R., & Piper, J. (1973). An associative thesaurus of English and its computer analysis. *The Computer and Literary Studies*, 153–165.

Koizumi, R., & In'nami, Y. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, *40*(4), 554–564.

Kojima, M., & Yamashita, J. (2014). Reliability of lexical richness measures based on word lists in short second language productions. *System*, *42*, 23–33.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Berger, C. M. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3),1030–1046.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*(2–3), 259–284.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal*, *75*(4), 440–448.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33.

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

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. Lund: CWK Gleerup.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208.

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift für Literaturwissenschaft und Linguistik*, *8*, 73–79.

MacWhinney, B. (2008). A unified model. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). New York, NY: Routledge.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17*(2), 457–472.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. *British Studies in Applied Linguistics*, *12*, 58–71.

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndsmill, UK: Palgrave Macmillan.

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459–488.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning*, *2*, 141–154.

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics*, *39*(6), 823–845. https://doi.org/10.1093/applin/amw050

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, *26*(1), 32–47.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *3*, 5–19.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins Publishing Company.

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary*. Heinle, Cengage Learning.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Nihalani, N. K. (1981). The quest for the L2 index of development. *RELC Journal*, *12*(2), 50–56.

O'Loughlin, K. (1995). Lexical density in candidate output on direct and semi-direct versions of an oral proficiency test. *Language Testing*, *12*(2), 217–237.

Paivio, A. (1971). *Imagery and verbal processes*. New York, NY: Holt, Rinchart, and Winston.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*(1p2), 1–1.

Pinchbeck, G. G. (2014, March). *Lexical frequency profiling of a large sample of Canadian high school diploma exam expository writing: L1 and L2 academic English*. Roundtable presented at the American Association for Applied Linguistics Annual Conference, Portland, OR. Retrieved from https://app.box.com/s/2772oofxkg6jyzmd3l7j

Pinchbeck, G. G. (2017). *Vocabulary use in academic-track high-school English literature diploma exam essay writing and its relationship to academic achievement*. Unpublished Ph.D. thesis, University of Calgary, UK.
Quirk, R., Greenbaum, S., Leech, G., & Svartik, J. (1985). *A comprehensive grammar of the English language*. London, UK and New York, NY: Longman.
Read, J. (2000). *Assessing vocabulary*. Cambridge, UK: Cambridge University Press.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke, UK: Palgrave McMillan.
Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499–520.
Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99–123.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, UK: Oxford University Press.
Ure, J. (1971). Lexical density and register differentiation. *Applications of Linguistics*, 443–452.
Webb, S. (2010). A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, *15*(4), 497–519.
West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans, Green and Co.
Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.
Yule, C. U. (1944). *The statistical study of literary vocabulary*. Cambridge, UK: Cambridge University Press.