CHAPTER 1

Defining and measuring lexical diversity

Scott Jarvis

Most existing measures of lexical diversity are either direct or indirect measures of the proportion of repeated words in a language sample, and they tend to be validated in accordance with how well they avoid sample-size effects and/or how strongly they correlate with measures of knowledge and proficiency. The present paper argues that such measures suffer from the lack of construct validity in two ways: (a) They are not grounded in an adequate or clearly articulated theoretical account of the nature of the construct of lexical diversity, and (b) they are not validated in relation to how well they measure lexical diversity itself, but rather in relation to how well they do or do not correlate with other constructs. The present paper proposes solutions to both of these problems by defining lexical diversity as a perception-based phenomenon with six measurable properties, and by calibrating the six objective properties against human judgments of lexical diversity. The purpose of the empirical portion of the paper is to determine how well a statistical model constructed on the basis of the proposed six objective properties is able to account for nine human raters' judgments of the lexical diversity found in 50 narratives written by adolescent learners and native speakers of English. The results support the proposed six-dimensional construct of lexical diversity, but also suggest the need for further refinements to how the six properties are measured.

1. Introduction

One of the core principles of language assessment – and indeed of psychometrics, testing, and research in general – is that all tests, measures, and indices need to be firmly grounded in a solid understanding of what is being measured (e.g., Bachman, 1990). That understanding is referred to as a construct, and the way it is articulated is referred to as a construct definition. According to Bachman and Palmer (1996), "we can consider a *construct* to be the specific definition of an ability that provides the basis for a given test or test task and for interpreting scores derived from this task" (p. 21, emphasis in the original). The validity, reliability, interpretability, comprehensiveness, and precision of a measure fundamentally require that the measure be derived from a construct definition rather than the other way around. The

necessary sequence of steps for developing a test or other type of measure is the following: "(1) identifying and defining the construct theoretically, (2) defining the construct operationally, and (3) establishing procedures for quantifying observations" (Bachman, 1990, p. 108). The sequence thus begins with a clear, specific, and well-articulated understanding of the underlying components of the phenomenon under investigation. This is then augmented in the second step with a description of the observable and measurable properties of the phenomenon. If the construct definition is adequate, it will indicate what those properties are and suggest the conditions under which they will be observable and can be measured (Bachman, p. 43). The third and final step in the sequence "is to establish procedures for quantifying or scaling our observations of performance" (p. 44).

More simply, these three steps can be understood as involving an interrelationship between the construct, its observable properties, and the measures that are developed to quantify those properties. The construct and its properties are both included in the construct definition, which is then used as the basis for developing a measure. Because there are often many alternative ways of interpreting and operationalizing a construct definition, a single construct definition can give rise to multiple measures. This may create confusion regarding which measure best captures the construct, but it also provides important feedback about which areas of the construct definition require further clarification and elaboration. Multiple measures derived from the same construct definition are not particularly worrisome because they are part of a natural cycle of discovery and development wherein theoretical advances give rise to improved forms of measurement, which in turn lead to further theoretical refinements. More worrisome, however, are measures that have been developed prior to or in the absence of an adequate theoretical construct definition, as well as measures that are used in ways that are incompatible with or reflect a poor understanding of the construct definition (assuming that there is a construct definition in the first place). In research dealing with the assessment of lexical diversity, such measures appear to be abundant. In this paper, I attempt to address this problem by proposing a fairly elaborate construct definition of lexical diversity and its observable properties. I also offer an early attempt at operationalizing the construct definition into a set of measures that tap into each of its theoretically proposed properties. Additionally, I conduct a preliminary validation of these measures through an analysis of how well – in concert with one another – they account for the holistic lexical diversity judgments of human raters.

2. Background and terminology

Although I assume that most readers will already be familiar with the terms *type*, *token*, *lexical diversity*, *lexical richness*, *vocabulary size*, and *vocabulary depth*, I

define them here briefly in order to avoid confusion later on. The terms *type* and *token*, when applied to vocabulary, refer to different ways of quantifying word frequencies in a language sample. To illustrate, the present sentence is 20 words long, so it can be said to consist of 20 lexical tokens. However, three of the 20 tokens are repetitions of words that appeared earlier in the sentence, either in exactly the same form (i.e., *to* and *to, 20* and *20*) or in a different form (*is* and *be*). (Here, I am not making a distinction between the copula and auxiliary *be*.) The separate occurrences of *to* and *to* count as two different tokens, but as only a single type. The same is true of the occurrences of *20* and *20*, as well as *is* and *be* (though some researchers prefer to analyze different forms of the same word as separate types). This means that the sentence in question consists of 20 tokens but only 17 types.

An imbalance between the number of types and tokens is the result of repetition, which is generally considered to be the opposite of lexical diversity. Thus, the more repetition in a text, the less lexically diverse it is – and vice versa. The term *lexical diversity* (or *diversity of vocabulary*, Carroll 1938) is used more or less interchangeably with the terms *lexical variation*, *lexical variety*, *lexical variability*, and *lexical flexibility* (see, e.g., Engber, 1995; Johnson, 1944; Read, 2000). In addition to mirroring the repetition rate, the patterns of behavior associated with these terms are believed to reflect a person's vocabulary size, or the number of words a person has in his or her active vocabulary (i.e., productive mental lexicon). The term *lexical richness* (or *vocabulary richness*, Yule, 1944) originally carried this latter meaning, referring to the "wealth of words at [the author's] command" (Yule, p. 83) – i.e., in his or her mental lexicon. More recently, the term *lexical richness* has been used as a cover term to describe a whole range of lexical measures assumed to reflect not just vocabulary size but also vocabulary depth, where *depth* refers to how well a person knows a particular word (see, e.g., Nation, 2001). The current meaning of *lexical richness* thus applies broadly to everything from lexical diversity through lexical sophistication (or a person's command of less-common words), to lexical density (or the proportion of content words in a text), and beyond (see, e.g., Engber, 1995; Read, 2000). One of the purposes of the present paper is to argue that *lexical diversity* is in fact not a subset of *lexical richness*. I will say more about this shortly, after first providing some additional historical context concerning how the field has arrived at its current understanding of lexical diversity.

In-depth discussions of the major historical developments that have taken place in the measurement of word frequencies are found, among other places, in Baayen (2001) and Malvern, Richards, Chipere, and Durán (2004). Malvern et al. describe a paper by Thomson and Thompson (1915) as being one of the first seminal papers in this area of research. Using sophisticated mathematical modeling, these researchers reasoned that the pattern of word repetitions found in a person's language use might provide a relatively precise indication of the number of words

in the person's active vocabulary. They recognized that some of their assumptions were simplistic and that word choice is governed by factors far beyond what they had accounted for, but they nevertheless held out hope that this problem could eventually be solved. One part of the problem is the fact that different words within a person's active vocabulary have different probabilities of being used. These probabilities are of course affected by syntactic constraints, semantic context, and perhaps some additional, more general principles. In a highly influential book, the American linguist George Kingsley Zipf (1935) observed a general principle that seemed to apply relatively consistently to the distribution of word frequencies. He noted that the most frequent word is usually roughly twice as frequent as the second most frequent word, and roughly three times as frequent as the third most frequent word, and so forth. This observation has come to be known as Zipf's law, a type of power law that has been found to apply to many phenomena both inside and outside of language (e.g., Clauset, Shalizi, & Newman, 2009). It involves a purportedly constant relationship between a word's frequency and its frequency rank, which can also be converted to a predicted relationship between a word and its repetition rate (Zipf, 1937). For present purposes, what is perhaps most noteworthy is that this relationship seemed to offer a solution to the problem of determining how to weight individual words – a solution that has continued to be adapted to work on the relationship between word use and vocabulary knowledge (e.g., Edwards & Collins, 2011, this volume; Ferrer i Concho & Gavaldà, 2009; Tuldava, 1996).

Zipf's law has not been without its detractors, however. Early work by Carroll (1938) and Chotlos (1944), for example, showed that Zipf's law does not accurately capture the relationship between word frequencies and ranks for the most frequent 20–30 words in a sample, and it furthermore produces varying levels of goodness of fit depending on the size of the sample being measured. (This latter shortcoming has been rigorously confirmed in the recent work of Baayen, 2001.) The sample-size problem was thus recognized early on, and it led to the search for a measure of word frequencies that would remain constant regardless of the length of the text being analyzed. Johnson (1939; 1944) – who first proposed the type-token ratio (TTR) as a measure of "vocabulary 'flexibility' or variability" (1944, p. 1) – was aware of this research and was also aware "of the tendency for the TTR to vary inversely with size of sample" (1944, p. 2). As a solution, he offered several different versions of TTR, including the "mean segmental TTR" (MSTTR), which involves splitting a text into several equally-sized segments, and using the mean TTR across all segments as the text's overall index of lexical variability.

Most other proposed solutions to the sample-size problem have been mathematically more sophisticated and have avoided segmenting texts into portions of a fixed length. The earliest such solution was proposed by Yule (1944), who offered

a formula that reflects the probability that any two words chosen randomly from a text will represent the same type. Higher probabilities mean a higher repetition rate – or less overall diversity. A functionally very similar but structurally simpler formula was later offered by Simpson (1949) as an index of diversity that reflects the concentration of groups (e.g., types) within a population (e.g., tokens). Still later work has attempted to find constants in the relationship between the total number of types, on the one hand, and the number of types occurring only once (e.g., Honoré, 1979), occurring twice (Michéa, 1971; Sichel, 1975), or occurring at multiple frequency levels (McKee, Malvern, & Richards, 2000; Sichel, 1986). Yet other work has attempted to correct the TTR so that it does not vary as a function of sample size (Carroll, 1964; Dugast, 1979; Guiraud, 1954; Herdan, 1960; Maas, 1972; Tuldava, 1993).

Despite the best efforts of the scholars who have proposed these measures, however, all have been found to vary as a function of sample size, to differing degrees (e.g., Baayen, 2001; Jarvis, 2002; Malvern et al., 2004; McCarthy & Jarvis, 2007; but see Baayen, 2001, pp. 29, 211 and Malvern et al. 2004, pp. 41–47). The only compelling exception to this that I have seen is the MTLD measure described and validated by McCarthy and Jarvis (2010; see also the chapters by McCarthy & Jarvis and Treffers-Daller in this book). My purpose in this chapter, however, is not to assess the value of different measures of lexical diversity in relation to their constancy across different sample sizes. Rather, I focus on the question of what it is that a measure of lexical diversity should be measuring in the first place. In other words, what is the nature of the construct? Ultimately, it is the answer to this question that will determine whether sample-size independence is really what we should be aiming for, and what other features of word use beyond or instead of the relationship between types and tokens we should be examining.

I acknowledge that theoretical construct definitions of lexical diversity are not completely lacking. Work dealing with the relationship between individuals' word use and the size of their active vocabularies – which began with Thomson and Thompson (1915) and continues in the work of Edwards and Collins (this volume) and a few other researchers – is based on clearly articulated principles of event probabilities, as well as on mathematical formalizations of observed tendencies (e.g., Zipf's law). According to Malvern et al. (2004), a mathematical model of lexical diversity that "stems from observing real behaviour" (p. 48) constitutes a valid theoretical construct definition. I do not disagree with this claim, but am nevertheless concerned about the fact that, for most existing measures of lexical diversity, the underlying construct definition is essentially just the equation that is used to calculate the index. While the existing equations do have roots in empirical observations, the practice of adopting a mathematical formalization of an observation as a yardstick for future observations is almost certainly not what

Bachman (1990) had in mind regarding the development of valid and maximally useful measures. Imagine, for example, if measures of language proficiency were developed on the basis of observations of error-frequency patterns in learners' language production rather than on the basis of a broader and deeper theoretical understanding of the nature of language, language use, language acquisition, and language proficiency. Clearly, this would result in inadequate measures of proficiency – even more inadequate than the ones currently in use. My concern is that this is essentially where we are with existing measures of lexical diversity. The following section describes a theoretical construct of lexical diversity that goes beyond the objective frequencies of types and tokens in a text.

3. Identifying the construct

Of all measures that have been used to assess the lexical diversity of samples of language use, TTR is perhaps the most intuitive and transparent. What could be more straightforward than the simple ratio of types to tokens? Nevertheless, TTR has come to be known as one of the least useful measures of lexical diversity because of the magnitude by which it varies in relation to sample size (e.g., Jarvis, 2002; Malvern et al., 2004; McCarthy & Jarvis, 2010). It is important to recognize, though, that the fault is not in the measure; the problem is the incompatibility between the measure and the construct. TTR and – I would suggest – all other existing measures of lexical diversity take as their input far too little information to account for the diversity of word use in a text. Regarding TTR and all measures that are calculated solely from type and token frequencies (e.g., Guiraud's index, Herdan's index), the problem is that they reduce an entire text to just two categories of words: (a) those that are novel and (b) those that recur. For convenience, I will refer to these two categories as first occurrences and repetitions.

Examples 1 and 2 represent texts whose words have been converted to 1s and 0s, where a 1 stands in place of a word that occurs for the first time in the text, whereas a 0 is a repetition of a word that occurred earlier in the text. Example 1 consists of 45 1s and 55 0s, for a total token count of 100. The TTR for this text (i.e., the number of 1s divided by the total token count) is thus 0.45. The second text consists of nine 1s and 11 0s, for a total token count of 20. The TTR for the second text is thus also 0.45. Both texts likewise have identical repetition indices (0.55), which are the exact opposites of TTR. These numbers are precisely as they should be, and any observer can confirm that the proportion of 1s and 0s in both examples is exactly the same. (Note: The distribution of 1s and 0s is not exactly the same in both texts, but TTR and most other measures of lexical diversity do not take ordering into account.)

Example 1:

11111111110001111100000000110110001101110000000000011101011010001101000001010100101110011001000010000

Example 2:

11111111000100000000

Intuition thus confirms that TTR is indeed an accurate index of the proportions of first occurrences and repetitions in texts regardless of how long those texts are. More strongly stated, TTR is an objective measure that is perfectly precise in its measurement of the phenomenon that it actually does measure. I will argue, however, that it does not measure lexical diversity. This can be seen quite clearly when we examine texts consisting of actual words. Example 3 is the original, non-binary version of Example 1, and Example 4 is the original, non-binary version of Example 2. Like Example 1, Example 3 consists of 45 types (where types are treated as lemmas – e.g., *is* and *are* are both counted as instances of *be*), 55 repeated words, and 100 tokens. Like Example 2, Example 4 consists of nine types, 11 repetitions, and 20 tokens. As before, the TTR and repetition values for both texts are identical (0.45 and 0.55, respectively), but this time the shorter text is perceptibly more repetitive than the first. Another way of saying this is that the shorter text has an unmistakably higher level of redundancy than the longer text.

Example 3:

> There are one hundred words in this paragraph. A token is a word occurrence, so the number of tokens in this paragraph is one hundred. A type, on the other hand, is a word treated as a category rather than as an occurrence. There are one hundred tokens in this paragraph, but how many of them are really different words – or different types? A simple way of counting the number of types is to count only the first occurrence of each word. When we do this, we find that there are forty-five word categories – or lexical types – in this paragraph.

Example 4:

> This clause has three nouns and a verb, and this clause also has three nouns and also has a verb.

The relationship (and contrast) between repetition and redundancy is fundamental to the present discussion. The repetition of a word that has occurred earlier in the text is often made necessary by grammatical or pragmatic constraints, but repetitions that are not warranted by such constraints are generally perceived as (unnecessarily) redundant (e.g., Bazzanella, 2011). The second clause in Example 4 is a stark illustration of redundancy: The only words in the second clause whose

repetition is contextually warranted are *and* and *this*. A less redundant wording of the second clause would be "and this one does, too" (although this would have a negative effect on the truth value of the clause). The word repetition rate is therefore an unreliable indicator of the amount of actual redundancy in the text, and this fact becomes even more problematic when one realizes the amount of repetition and redundancy that occurs in the form of synonyms and paraphrases (e.g., Reynolds, 1995). Critically, repetition in its purest sense is an objective phenomenon, whereas redundancy is fundamentally subjective – not in the sense of being a matter of personal taste and thus varying from one individual to the next, but in the sense of being grounded in human perception. At its most basic level, redundancy involves the perception of excessive or unnecessary repetition.

The notion that redundancy is a subjective (perception-based) construct does not necessarily mean that it cannot be measured objectively, but it does mean that it cannot be measured accurately through objective means until the researcher fully understands all of the factors that affect the way it is perceived, and not until the researcher also understands how to apply proper weights to each of those factors. This is of course true regarding the objective measurement of other subjective constructs, too, such as language proficiency and color (e.g., Goldstein, 2007). I will return to this point shortly. The crucial point for now is that a measure of repetition will never suffice as an adequate index of redundancy because redundancy is a far more complex phenomenon than repetition. I would make the same claim regarding lexical diversity – that it is fundamentally a subjective phenomenon that is far more complex than the more purely objective phenomenon that TTR and all other existing indices in its class are designed to measure. More specifically, I am claiming that all existing measures of lexical diversity are actually not measures of the complex, perception-based construct that I am referring to here as lexical diversity, but are rather measures of a simpler, objective construct of what should perhaps be called lexical variability. Johnson (1944) himself used the term *vocabulary variability* when introducing TTR and describing what it and some of its variants measure. This seems quite suitable because *variability* carries connotations of an objective phenomenon – a phenomenon that is the way it is regardless of how a human observer might perceive it. *Diversity*, on the other hand, is largely a matter of perception, as I will explain shortly.

To take stock, lexical repetition and lexical variability are objective constructs, and they are the mirror opposites of each other within the objective realm (see Table 1). On another dimension, lexical repetition is related to lexical redundancy, with the former being an objective construct and the latter a subjective construct. My proposal here is that the same relationship exists between lexical variability and lexical diversity. Crucially, this means that measures of variability are inadequate measures of diversity just as measures of repetition are inadequate measures

Table 1. Taxonomy of constructs related to the objective-subjective and novelty-recurrence dichotomies

	Objective	**Subjective**
Novelty	Lexical variability	Lexical diversity
Recurrence	Lexical repetition	Lexical redundancy

of redundancy. As I described earlier, lexical diversity and lexical redundancy are highly complex phenomena that are grounded in human perception and are subject to potentially numerous influences beyond those that affect variability and repetition.

Before going further, I should acknowledge that the recognition of lexical diversity as a subjective construct does not solve the sample-size problem that has plagued nearly all measures of what I am referring to here as lexical variability. It does, however, allow us to see the problem in a new light. Lexical measures designed to negate the effects of text length on lexical variability may be quite useful, as McCarthy and Jarvis (2007; 2010; this volume) have argued, but until researchers understand the specific effects of text length on the perception of lexical diversity, such measures might not actually improve the precision of what we are really trying to measure. As I alluded to earlier, it appears that the perception of redundancy – and, by extension, also lexical diversity – changes as a text grows longer, in which case negating the effects of text length could be counterproductive.

As I also mentioned earlier, the characterization of lexical diversity as a subjective construct does not mean that it can only be measured through subjective human judgments. Color research is a good model here. Colors do not have an existence independent of the way they are perceived. However, just because color perception is subjective does not mean that everyone sees colors differently. Instead, there is a great deal of inter-subjective consistency in how colors are perceived, and this is largely due to the fact that structures in the human retina react in very specific ways to certain wavelengths of light. Careful investigations of this relationship have ultimately made it possible to predict with high levels of accuracy how particular concentrations of electromagnetic energy (i.e., light) within a given stimulus will be perceived by human judges in terms of hue, saturation, and brightness (e.g., Goldstein, 2007). This does not mean that speakers of all languages necessarily divide the color spectrum in the same way (see, e.g., Athanasopoulos, 2009), but it does mean that there is a straightforward and predictable relationship between wavelengths of visible light and human perceptions of the areas of the color spectrum they represent. Crucially, because of early studies that determined the relevant dimensions of color perception and then calibrated objective measurements of those dimensions with subjective

human perceptions of colors, we can now rely solely on objective measurements to produce or predict specific color effects (e.g., the ability to mix different colors of paint in precise proportions in order to produce a specific color effect).

A similar goal may be achievable in language research, too. In fact, a good deal of progress toward such a goal can already be seen in the measurement of language proficiency (e.g., Fulcher & Davidson, 2007) and even in the more specific measurement of lexical proficiency. Regarding the latter, recent work by Crossley, Salsbury, McNamara, and Jarvis (2011a; 2011b; see also Crossley, Salsbury, and McNamara, this volume) has shown, first of all, that human raters with relatively minimal training display a high level of inter-rater consistency in their judgments of learners' lexical proficiency. Second, our research has shown that human raters' lexical proficiency judgments can be predicted with up to 60% accuracy on the basis of a properly weighted and carefully selected combination of just four objective lexical measures (lexical variability, word imageability, word familiarity, and word hypernymy). These results give me confidence that theoretically motivated and carefully developed objective measures might be able to predict human judgments of lexical diversity, as well.

4. Defining the construct

In order to arrive at a satisfactory measure of lexical diversity as a subjective construct, we need to (a) determine the dimensions of lexical diversity, (b) devise valid measures of those dimensions, and (c) combine them in such a way that we calibrate them with actual human perceptions. In the remainder of this chapter, I will describe some of the progress I have made in relation to these three steps. I concentrate in this section on the first step, which involves identifying the internal factors – internal dimensions, components, or properties of the construct – that determine how lexical diversity is perceived.

Variability. Just as repetition is known to be an inherent property of redundancy, so too must variability be an inherent property of diversity. Variability (or what I have referred to in previous work as variegation; Jarvis, 2012) is thus postulated as the first property of lexical diversity.

Volume. The second property is sample size, or volume (to use a shorter and simpler term). From the research reviewed earlier, there is reason to believe that the lexical variability of a text changes as the text grows longer. With TTR and measures that take as their input only types and tokens, variability is usually found to decrease with increasing text length; with probability-of-occurrence measures, on the other hand, variability is usually seen to increase as the text grows longer (see, e.g., McCarthy & Jarvis, 2007). Volume (i.e., text length) thus has clear effects on almost

all measures of variability. It also appears to have clear effects on how lexical diversity is perceived (cf. Turlik, 2008), as discussed in the following section.

Evenness. Beyond variability and volume, the third postulated property of diversity is evenness (or what I have referred to in previous work as balance; Jarvis, 2012). Evenness refers to how evenly the different words in a text are represented. Another way of saying this is that it refers to how evenly the tokens in a text are distributed across types. By way of illustration, Figure 1 shows the frequency distribution of the 45 types in the 100-word text from Example 3. The first bar in Figure 1 shows that 24 of those 45 types are words such as *when* that occur only once in the text. Seven types are words such as *category* that occur twice in the text, and so on. The most frequent word in the text is the verb *be*, which in various forms occurs 8 times in the text. As this chart shows, there was only one word with 8 occurrences.

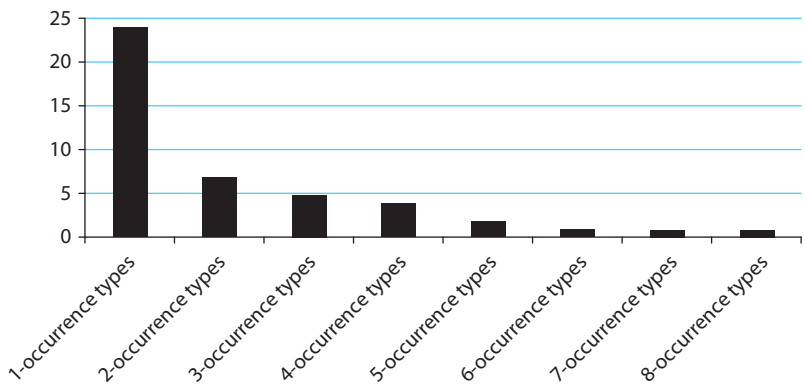

Figure 1. Frequency distribution of the lexical types in the 100-word text in Example 3

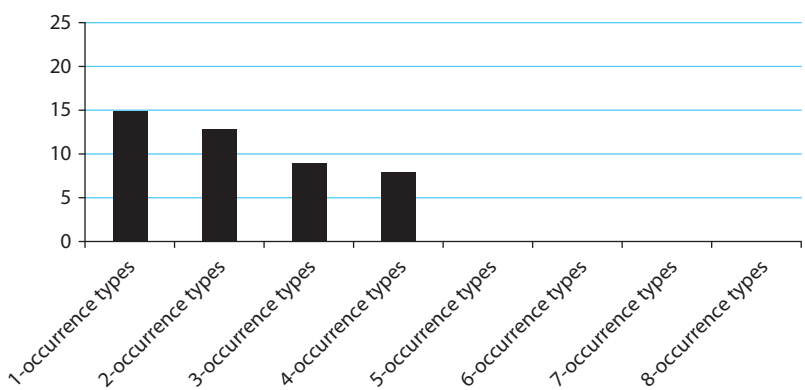

Figure 2. Frequency distribution of a hypothetical 100-word text

Compare Figure 1 with Figure 2, which represents a hypothetical text having exactly the same number of types and tokens as the text represented by Figure 1. Both texts have 45 types and 100 tokens, so they both have a TTR of 0.45 and a repetition index of 0.55. The difference between the two texts is that the tokens are distributed differently across types. In the text represented in Figure 2, there are no types with more than four tokens, and there is also less difference between the number of types that occur once versus twice, and so forth. Most importantly, there is a smaller range and standard deviation of tokens per type in Figure 2. In other words, the types in Figure 2 are more evenly balanced than the types in Figure 1. Now, of course, any actual text having the profile shown in Figure 2 would probably sound very strange (in English, at least), but the main point for now is that, even with the same number of types and tokens, two texts do not necessarily reflect the same level of lexical diversity or even lexical variability. Although TTR and its derivatives do not show any difference between these two texts, more sophisticated measures – which are based on probabilities associated with the frequency spectrum of a text – do. For example, Yule's K (Yule, 1944), Simpson's D (Simpson, 1949), the D measure produced by *vocd* (Malvern et al., 2004), and a corresponding measure that makes use of the hypergeometric distribution (McCarthy & Jarvis, 2007), all show that the text represented in Figure 2 has a higher level of lexical variability than the text represented in Figure 1. Although it is possible that this is merely an objective, statistical outcome that has no impact on the perception of lexical diversity, I have included evenness as a possible property of lexical diversity worthy of further examination.

Rarity. A fourth potential property of the construct is rarity, or the use of less common, less frequent words. Although no one, to my knowledge, has yet operationalized rarity into a measure of lexical diversity, rarity (aka lexical sophistication or the lexical frequency profile) has received a good deal of attention in the literature as one of many loosely related measures of lexical richness that may be indicative of a learner's vocabulary knowledge or even level of language proficiency more generally (e.g., Arnaud & Savignon, 1997; Laufer & Nation, 1995; Linnarud, 1986; Meara & Bell, 2001; Read, 2000; see also Edwards & Collins, 2011; this volume). However, if it turns out that rarity works together with variability, volume, and evenness in determining how lexical diversity is perceived, then this may show that lexical richness is not just a grab bag of loosely related lexical measures after all. Instead, a good number of these measures may be directly linked to properties of a unitary construct.

Dispersion. As I have recently discovered, the field of ecology already has a complex view of diversity similar to the one I am proposing here for lexical diversity. Drawing from ecologists' insights concerning the properties of biodiversity therefore seems appropriate at this early stage of developing a theoretical

understanding of diversity in our own field. One of the aspects of diversity that has received a good deal of attention in the field of ecology is the notion of dispersion, which refers to the spatial distribution of species in relation to one another (e.g., Walker, 2011). The core of this notion has to do with the degree to which tokens of each type are dispersed evenly throughout a domain (e.g., text) as opposed to being clustered in close proximity to other tokens of the same type. As demonstrated earlier in Example 4, the close proximity of tokens of the same type increases the perception of redundancy. Accordingly, we can assume that greater dispersion will lead to higher levels of perceived lexical diversity. That is, we can expect that a text whose tokens of the same type are dispersed far apart from one another will be perceived as being more lexically diverse than a comparable text whose tokens of the same type are more closely clustered together.

Disparity. Yet one more property of diversity we can glean from the field of ecology is disparity, or the degree of differentiation between species within an ecosystem (e.g., Barraclough, Hogan, & Vogler, 1999). Gould (1990) describes it like this: "Three blind mice of differing species do not make a diverse fauna, but an elephant, a tree, and an ant do – even though each assemblage contains just three species" (p. 49). Translating this to the realm of lexis, disparity involves the degree of differentiation between lexical types in a text, and this relates closely to the points raised by Reynolds (1995) concerning the fact that, on a semantic level, repetition and also redundancy can arise through the use of synonyms and paraphrase – not just through the literal repetition of the same precise forms. Even on a formal level, some words are more similar to each other than other words are. Both types of lexical disparity – formal and semantic – could potentially have effects on the perception of lexical diversity. I will return to this issue shortly.

To take stock, the six postulated properties of lexical diversity are variability, volume, evenness, rarity, dispersion, and disparity. Many of these properties are already recognized as aspects of diversity in other fields, such as ecology, where there already exist indices of component properties such as volume, evenness, dispersion, and disparity (e.g., Barraclough et al., 1999; Ginebra & Puig, 2010). The indices used in ecology and other fields are not necessarily more mathematically sophisticated than those already in use by linguists (see, e.g., Baayen, 2001; Malvern et al., 2004), but they do seem to be better tailored to their intended purposes (see, e.g., Chao & Jost, in press). Consequently, the range of solutions ecology and other fields have found for measuring various dimensions of diversity will likely serve as useful models for future work on the measurement of lexical diversity in our field (see Jarvis, 2013). Future work might likewise uncover additional, relevant properties of diversity that I have not yet accounted for, but the six already introduced reflect my current understanding of the full range of properties that are likely to affect the perception of diversity in general, and lexical diversity in particular.

As a starting point for investigating the potential effects of these properties on the perception of lexical diversity, I created two tasks where human judges were presented with a series of short language samples that they were asked to compare and then decide which samples represented higher levels of lexical diversity. Lexical diversity was defined for them as "the variety of word use that can be found in a person's speech or writing." The first task included six pairs of sentences that are relevant to the present paper; in each pair of sentences, just one property was manipulated at a time (see Appendix A), although dispersion was not manipulated at all in any of the pairs of sentences, whereas disparity was manipulated in two sets of sentences – once in relation to formal disparity (item 4) and once in relation to semantic disparity (item 5). This task was administered to 130 participants, 109 of whom were native English speakers (98 undergraduates, 6 graduates, 5 other) and 21 of whom were nonnative English speakers (8 undergraduates, 12 graduates, 1 other). Their judgments showed a clear and significant effect for variability, volume, rarity, and evenness, but no significant effect for either semantic or formal disparity. However, the effects of rarity may have confounded the effects of semantic disparity in item 5 (see Appendix A), given that the words *honest*, *truthful*, and *sincere* are less frequent in English (mean rank of 7,733 in the American National Corpus) than the words *smart*, *energetic*, and *pretty* (mean rank = 4,112). When I later replaced *smart* with *intelligent* (rank = 6,664) and *pretty* with *beautiful* (rank = 1,164) and administered this pair of sentences to 22 new participants, the expected significant effect was found. The statistical results for each property are shown in Table 2 in order of decreasing effect size. In this table, disparity refers to semantic disparity and not formal disparity, and the results for disparity in Table 2 represent the new item administered to the latter group of 22 participants. As one can see in the table, variability and volume show strong effects on the participants' judgments of lexical diversity. Disparity, rarity, and evenness have moderate effects, whereas the effects of dispersion were not tested.

Table 2. Effect sizes of five properties on the perception of lexical diversity in the paired-sentences task

Property	X^2	df	p	Effect size
Variability	73.50	1	< .001	Phi = .88
Volume	70.04	1	< .001	Phi = .85
Disparity	6.23	1	.01	Phi = .69
Rarity	31.12	1	< .001	Phi = .68
Evenness	23.28	1	< .001	Phi = .52

The second task was a paragraph-sorting task involving six paragraphs that describe the same scene in the Chaplin film *Modern Times*. One of the paragraphs was written as a baseline text – as a typical example of how a native speaker would be likely to describe the scene. The remaining five paragraphs represent modifications of the baseline in such a way that one property was increased or decreased (see Appendix B). The task was administered to 38 participants, all but one of whom were students (undergraduate and graduate, both native and nonnative English speakers) at an American university. The participants were given the paragraphs and asked to sort them in the order from most lexically diverse to least lexically diverse. The participants were given the same definition of lexical diversity as the previous participants. The participants' mean rankings and standard deviations are given in Table 3. The table shows that the text with high rarity (i.e., less common words, such as *destitute* instead of *poor*, as measured in relation to their frequency ranks in the American National Corpus) was ranked as the most lexically diverse text. The paragraph that was the longest was judged to be the second most lexically diverse, and the paragraph with the highest number of lexical types was judged to be the third most lexically diverse. The effects of evenness and disparity are a little more difficult to evaluate. The fact that these two texts – which were modified to have low levels of a particular property – were ranked as less diverse than the ones with high levels of a specific property is fully in line with expectations. On the other hand, the fact that they were ranked higher than the baseline was not expected. The fact that they were modified at all, however, may have created a certain novelty effect that made them seem more diverse (i.e., more varied, less predictable) than the baseline text. In any event, these results, together with the results of the first task, suggest that all five properties may indeed have some effect on the perception of lexical diversity. The properties with the strongest effects appear to be variability, volume, and rarity, though not necessarily in that order. Further work will be necessary to determine the exact weights of each of these properties, and also to determine the potential effects of dispersion. I address these issues to some degree in the following section. In the meantime, I conclude

Table 3. Mean ranks for paragraphs manipulated in relation to specific properties

Paragraph	Mean rank	Standard Deviation
High rarity	1.34	0.53
High volume	2.24	1.30
High variability	3.26	1.25
Low evenness	4.13	0.74
Low disparity	4.79	1.02
Baseline	5.24	1.05

this section by asserting that the theoretical tenets, logical reasoning, informal observations, and empirical evidence hitherto presented do indeed appear to validate a construct definition of lexical diversity that rests on these six properties.

5. Operationalizing the construct and calibrating the measures

At the beginning of the preceding section, I suggested that the development of a satisfactory measure of lexical diversity requires (a) determining the inherent properties of lexical diversity, (b) devising valid measures of those properties, and (c) combining and calibrating these measures with human judgments of lexical diversity. I addressed the first of these steps in the previous section, and will address the remaining two in this section.

Variability. The second step, stated differently, involves the search for or development of accurate, precise, and pure measures of each property. This is quite a challenge because most existing lexical measures are not pure measures. As already mentioned, for example, measures of variability tend to be affected by volume and evenness. There is only one measure of variability I am aware of that does not vary as a function of volume or evenness, and this is the MTLD measure developed by McCarthy (McCarthy, 2005; McCarthy & Jarvis, 2010; this volume; but see Covington & McFall, 2010; Johnson, 1944). MTLD is consequently the measure I tentatively propose be used as a measure of the property of variability within the larger construct of lexical diversity. MTLD is calculated by first identifying within a text the maximum number of running words whose TTR value remains above a certain threshold (e.g., 0.71). Once a running string of words crosses the TTR threshold, the program records the length of the string minus one (i.e., the length of the string just before it crossed the threshold), and then begins looking for the next maximum string of words that remains above the TTR threshold. The program continues to do this until it reaches the end of the text, at which point it identifies such sequences again in the reverse direction. MTLD takes as its final value the mean length of all such word sequences that remain above the TTR cutoff criterion. MTLD is obviously not a very complex measure, but it clearly does measure lexical variability, and does so in a way that is evidently not affected by volume, evenness, or any of the other component properties of lexical diversity.

Volume. Although most existing measures of lexical diversity are designed to minimize the effects of volume, the research described earlier provides compelling support for the notion that volume actually contributes to the perceived lexical diversity of a language sample. The simplest measure of volume is the number of words (tokens) in a text, and for the purposes of the present study, I will assume that the simplest solution is the best solution until evidence to the contrary

suggests otherwise. I therefore adopt the number of word tokens in a text as my tentative measure for volume. Perhaps the most substantial challenge in applying this measure is defining what a word is. The assumption that a word is simply what our spelling conventions determine it to be ultimately will not suffice; we need more principled ways of determining what a lexical item is that extends to compound words and, in some cases, perhaps also to whole phrases (e.g., Nation, 2001; Meunier & Granger, 2008). We also, of course, need a measure that works cross-linguistically.

Evenness. Next, the simplest way to measure evenness is perhaps to use the standard deviation of the number of tokens per type in a text. This would provide an indication of the range and magnitude of differences in the number of tokens found for each type. There may be some disadvantages to doing this, however, because the standard deviation might be affected by volume. Ginebra and Puig (2010) offer a series of alternative solutions involving what they describe as mixed Poisson models for measuring evenness. They acknowledge, though, that even these complex models have shortcomings and are subject to fluctuations depending on the nature of the sample. It might ultimately be worthwhile to follow their recommendations, but as a first attempt, I will adopt the simpler standard-deviation solution as a tentative measure of evenness.

Rarity. The most straightforward way to measure rarity is to assess the overall commonness of the words used in a text in relation to how frequently those words occur in the language in general. This of course requires the use of a large and well-balanced reference corpus, such as the British National Corpus (BNC; www.natcorp.ox.ac.uk), the American National Corpus (ANC; americannationalcorpus.org), or the Corpus of Contemporary American English (COCA; corpus.byu.edu/coca). The COCA is the largest of these, consisting of over 400 million words, but for now I have chosen to use the BNC (100 million words) as my reference corpus for texts written in English due to the fact that the rank-ordered lemma list available for the BNC is larger than the equivalent list available for the COCA. I could not find a rank-ordered lemma list for the ANC (over 22 million words), but the resources available from the ANC website do include a rank-ordered list of the nearly 30,000 unique lexemes that occur in the ANC, and this list could probably be transformed into a rank-ordered list of lemmas without too much difficulty. I may attempt to do this in the future, but in the meantime, I will use the BNC lemma list as the basis for my rarity measure.

Choosing a reference corpus and finding a suitable rank-ordered list of lemmas from that corpus is only the beginning. There is still the question of what to do with this information. Perhaps the most straightforward measure of rarity would entail simply identifying each word in the data with its lemmatized rank order in the reference corpus, and then using the mean rank for all words in the

data as the text's index of rarity. This is what I will do as a first attempt, although I acknowledge that there might be good reasons to convert the rank orders to frequency bands, and perhaps also to calculate the index of rarity as the proportion of words within a particular band rather than as a mean rank (e.g., Laufer & Nation, 1995). Clearly, a good deal of exploratory work will be needed to arrive at an optimal measure of rarity and all other properties of lexical diversity.

Dispersion. Regarding the measurement of dispersion, various indices of dispersion can be found in different disciplines. In statistics, a dispersion index is calculated as the ratio of the variance to the mean (Upton & Cook, 2006), which gives an indication of how tightly clustered a set of values is around the mean. An index of dispersion can also be found in the literature dealing with lexical analysis. Gries (2009), for example, describes a measure called Juilland's *D*, which is similar to a standard deviation in that it indicates how evenly represented a particular word is across different parts of a text. The problem, though, is that this measure is applied to individual words separately rather than providing an overall index of the lexical dispersion in the text as a whole. The simplest way of calculating a dispersion index for the text as a whole would probably be to calculate the mean distance between different tokens of the same type, and to aggregate this value for all types in a text. There may be some unforeseen negative consequences of doing this, but it seems prudent to start with this as the simplest solution.

Disparity. Finally, disparity – or, more specifically, semantic disparity – might be the most difficult property to measure, for three reasons. The first reason is that there are so many different levels on which separate words can be semantically related. They can, for example, be related through synonymy and antonymy, hypernymy, frequent proximal co-occurrence, or various types of mental association (e.g., Landauer, McNamara, Dennis, & Kintsch, 2007; Meara, 2009; O'Grady, Archibald, Aronoff, & Rees-Miller, 2010). The second reason is that it is not clear how to rate the degree to which words are related to one another within and across these different levels. The third reason is that, even after the first two problems are solved, it will take a great deal of effort to create cross-referenced semantic-relationship tables that allow for the automated computation of the overall degree of semantic relatedness among the words in a text. In the meantime, two temporary solutions are available. The first is to use a Latent Semantic Analysis (LSA) measure available in Coh-Metrix (Landauer et al., 2007), and the second is to create a measure that uses the WordNet (wordnet.princeton.edu) semantic sense index to determine the mean number of words in a text that share the same semantic sense. In the present paper, I will opt for the latter solution as it has a more straightforward interpretation and is computationally simpler.

A list of the six properties of lexical diversity and the measures I have used to operationalize them is given in Table 4. Whether these particular measures are the

Table 4. The six properties of lexical diversity and the measures adopted to measure them

Property	Measure
Variability	MTLD
Volume	Total number of words in the text
Evenness	SD of tokens per type
Rarity	Mean BNC rank
Dispersion	Mean distance between tokens of a type
Disparity	Mean number of words per sense

best way to operationalize the properties is still to be determined, and this is indeed one of the purposes of the present study.

The third step in developing a satisfactory measure of lexical diversity involves combining and calibrating the proposed measures with what can be considered to be the most authoritative ratings of lexical diversity, which I have argued will necessarily involve human judgments. I will do this in relation to human judgments of both language proficiency and lexical diversity. The data in question involve written film descriptions of the eight-minute "Alone and Hungry" segment of Chaplin's film *Modern Times*. As described in Jarvis (2002) and McCarthy and Jarvis (Chapter 2 of this volume), the essays were written by 210 Finnish-speaking and Swedish-speaking learners of English living in Finland and enrolled in grades 5, 7, and 9. A breakdown of the learners by L1 background, grade, and years of English instruction is given in Table 5.

The learners' essays, which ranged in length from 24 to 578 words (mean = 218.89, sd = 102.81), were entered into a computer database without correcting any of their spelling, grammar, punctuation, or stylistic errors. The essays were later printed out and given to two trained raters at Indiana University to be rated for writing quality. The raters used a rating scale that corresponded to the proficiency levels in Indiana University's Intensive English Program (IEP). There are seven levels in the program, and the raters used holistic ratings of 1–7 to indicate

Table 5. Breakdown of participant groups

Group	*n*	L1	Grade	English	Swedish	Finnish
F5	35	Finnish	5	2 yrs	0 yrs	Native
F7	35	Finnish	7	4 yrs	0 yrs	Native
F9a	35	Finnish	9	6 yrs	2 yrs	Native
F9b	35	Finnish	9	2 yrs	6 yrs	Native
S7	35	Swedish	7	2 yrs	Native	4 yrs
S9	35	Swedish	9	4 yrs	Native	6 yrs

which level a particular essay would be placed into. They also used plusses and minuses to make finer distinctions, and used 0 and 0+ for any essays that reflected writing quality below that expected in the lowest level of the IEP, as well as 8-, 8, and 8+ for any essays that reflected writing quality beyond the highest level of the IEP. All of these ratings were then converted to a 26-point scale for computational convenience (0 became 1, 0+ became 2, 1- became 3, ... 8+ became 26). The inter-rater reliability for the two raters was $r = 0.94$ ($p < .001$), and I used the mean of their ratings as the proficiency score for each text.

To calculate indices for each of the proposed six properties of lexical diversity, I first went through the data and lemmatized all words into their base forms (e.g., stole > steal; steals > steal; steal > steal). I then used the programming language *Perl* to create scripts that would count types and tokens for each text, match types in the data with their ranks in the BNC lemma file and with their semantic senses in the WordNet sense file, and perform all of the calculations described earlier in order to produce the measures listed in Table 4.

The correlation matrix in Table 6 shows the relationship among the proficiency ratings and all six measures for the 210 texts in question. It is quite interesting that five of the six diversity measures are significant predictors of learners' writing proficiency, and it is particularly interesting that dispersion turned out to be the strongest predictor. An equally interesting and surprising result is that that rarity turned out not to be a significant predictor of proficiency. This is surprising because rarity should be expected to increase with the size of learners' vocabulary knowledge, and vocabulary knowledge should increase with proficiency. However, the nature of the task might at least partially account for the lack of a significant correlation between rarity and proficiency in these results: Film-based and picture-based narratives do not lend themselves to a great deal of content variety. Although greater levels of rarity can be achieved in this type of task when writers choose less-frequent words over their more common synonyms (e.g., *destitute*

Table 6. Matrix of Pearson bivariate correlation coefficients

	PROFIC	VARIAB	VOLUME	EVENN	RARITY	DISPER	DISPAR
PROFICIENCY	1						
VARIABILITY	0.41**	1					
VOLUME	0.72**	0.24**	1				
EVENNESS	0.55**	–0.15*	0.87**	1			
RARITY	0.12	0.15*	0.17*	0.07	1		
DISPERSION	0.77**	0.44**	0.94**	0.73**	0.13	1	
DISPARITY	0.55**	0.34**	0.62**	0.49**	0.06	0.64**	1

*significant at $p < 0.05$, **significant at $p < 0.01$

versus *poor*, *collide* versus *crash*), doing so often results in less natural-sounding texts. Because rarity can be excessive and can also reflect a weak knowledge of word-choice constraints and conventions, we should perhaps not expect a strictly linear relationship between rarity and proficiency.

Regarding the correlations among the six properties themselves, we should expect them to correlate significantly with one another given that they are hypothesized to be measuring different aspects of the same construct, but it would be ideal if their correlations were only moderate so as to indicate that they are not varying as a function of each other. From this perspective, the correlations that are the most disappointing are the ones higher than 0.80 – namely, the correlations between volume, on the one hand, and evenness and dispersion, on the other. The high correlations found here seem to suggest that the measures I have adopted for evenness and dispersion vary with text length. If so, alternative measures will need to be found in the future.

What really matters, of course, is how well these measures work in concert with one another in predicting human judgments of lexical diversity – not proficiency per se. This is the issue that I address next. Unfortunately, I have not yet been able to have all 210 learner texts rated for lexical diversity, but I have had 37 of them rated, along with 13 additional film-prompted narrative texts produced by American English-speaking participants in grades 5, 7, and 9. The following analysis focuses on these texts. The 50 texts were chosen on the basis of their proficiency ratings. (The English speakers' texts were rated for proficiency using the same procedures and by the same raters mentioned earlier.) Recall that the texts were rated on a 26-point proficiency scale. The 50 texts I selected had been given ratings of 7.5 (n = 5), 10 (n = 7), 13 (n = 20), 16 (n = 10), and 20 (n = 8). These are the scores that had the highest numbers of texts associated with them, and were also relatively well distributed across the three L1 groups – except that scores of 7.5 could be found only for Finnish and Swedish speakers, and scores of 20 could be found only for English and Swedish speakers.

The 50 texts were rated by 11 participants. The first eight participants were ESL teachers with at least three years of teaching experience, and the final three were graduate students at Ohio University. All 11 participants were highly proficient speakers of English, although two of the ESL teachers were nonnative speakers of English. Each participant rated either 10 or 20 texts, which meant that each text was rated by at least two participants, and some were rated by three. In order to ensure that the participants' judgments were not affected by lexical errors found in the texts, spelling and grammar errors were corrected before the texts were given to the participants. In many cases, this required the addition or deletion of various function words, especially articles, and this of course has consequences for the number of words in each text as well as for the writing quality of the text.

Nevertheless, editing the texts in this way seemed necessary in order to avoid the potential influence of errors on the raters' perception of lexical diversity.

Whereas the proficiency ratings had been performed by trained raters with a great deal of rating experience and whose understanding of writing proficiency had been normed on the basis of clear benchmarks and rubrics, the participants recruited to perform the ratings of lexical diversity had never previously judged the lexical diversity of a text. Although it would have been possible to create a lexical diversity rubric to assist the raters in their judgments, this would have resulted in a severe circularity of purpose because the rubric would have reflected the six proposed properties, yet the purpose of the study in the first place was to determine whether these six factors affect human judges' perceptions of lexical diversity without their being told what to look for. Therefore, in order to determine whether human judges already have an intuitive sense of what lexical diversity is, and in order to determine whether their intuition is grounded in the six proposed properties, they were given the following minimal instructions:

> Please quickly read each essay through once. After completing each essay, rate its lexical diversity (defined simply as the variety of different words used) on a scale of 1 to 10. We are looking for your perception, so go with your first instinct. Provided as an example is one essay whose lexical diversity you may consider to be a five out of ten.

The essays were administered in sets of 10 essays, and each set of 10 was rated by either two or three raters. Although inter-rater reliability cannot be assessed with high levels of confidence in cases of only 10 pairs of ratings, this provides at least some indication of the degree of congruence between the raters. In most cases, the Pearson correlation coefficient was above 0.45, which is very low, but is understandable in the context of correlation tests involving only 10 texts at a time. However, two of the 11 raters were not consistent with the others, and they also showed low correlations between their ratings and all or nearly all six measures of diversity. Fortunately, the texts rated by either of these two raters were texts that had been rated by two other raters. Consequently, the ratings produced by the two outliers were completely removed, and the lexical diversity rating for each text was ultimately calculated as the mean of the two remaining raters' scores.

Because the 50 texts were edited before they were given to the raters, I re-calculated the indices for all six properties using the *Perl* scripts mentioned earlier in order to make sure that the indices reflected the actual texts that were rated. I then ran bivariate Pearson correlation tests on each pair of these indices, as well as between these indices and the participants' lexical diversity ratings. The results of these tests are shown in Table 7.

Table 7. Matrix of Pearson bivariate correlation coefficients

	DIVERS	VARIAB	VOLUM	EVENN	RARITY	DISPER	DISPAR
DIVERSITY RATING	1						
VARIABILITY	0.31*	1					
VOLUME	0.67**	0.21	1				
EVENNESS	0.53**	–0.17	0.89**	1			
RARITY	0.26	0.07	0.03	–0.03	1		
DISPERSION	0.64**	0.43	0.94**	0.74**	–0.01	1	
DISPARITY	0.46**	0.25	0.58**	0.48**	–0.09	0.56**	1

*significant at $p < 0.05$, **significant at $p < 0.01$

Regarding the relationship between the six indices and the raters' perceptions of lexical diversity, the first numeric column of Table 7 indicates quite clearly that these indices do indeed seem to predict the human judgments. The one exception is rarity, which is not significantly correlated with the diversity ratings, but this may be the result of the low N (= 50); the nonsignificant correlation between rarity and the diversity ratings is nearly as high as the significant correlation between variability and the diversity ratings. As before, we also see a potential problem in the correlations between volume, on the one hand, and evenness and dispersion, on the other. These high correlations suggest that the evenness and dispersion measures I have adopted capture a relationship that varies with text length. One additional potential problem is the correlation between disparity and the diversity ratings. Even though the value of 0.46 seems encouraging, the correlation is in the opposite direction from what my model of lexical diversity would have predicted. That is, the disparity measure used in the present study is essentially an index of the relative number of synonyms in a text. That means that the higher the index, the less semantic disparity there is between the word types in the text. My model would have predicted a negative correlation between this and the perception of lexical diversity. What may nevertheless be happening in the data is that synonyms are appearing in place of repetitions of the same word in a way that maintains coherence where necessary while increasing diversity on other dimensions (e.g., variability). In other words, the positive correlation between disparity and the diversity ratings is perfectly logical, but may suggest the need for another way to measure disparity.

One way to assess the validity of the measures proposed in this chapter is to analyze them in a multiple-regression model using the lexical diversity ratings as the dependent variable and the six indices as independent variables. Doing so with the current data produces a model having an adjusted R^2 of 0.48 ($F[6,43] = 8.55$, $p < 0.001$). Although significant, this is a somewhat disappointing result given that

the six indices together account for less than 50% of the variance in the diversity ratings. I suspect that the problem lies both in how the measures have been operationalized and in the fact that the human judges' lexical diversity ratings are not fully consistent with one another. Even though the ratings are consistent enough to show that the raters do indeed seem to have a similar general sense about what lexical diversity is, it is clear that perceptions of lexical diversity – just like perceptions of language proficiency – will ultimately need to be calibrated through rater training and norming (cf. Malvern & Richards, 2002). Again, though, the danger in doing this before the field is absolutely certain what lexical diversity is made up of, is that it would result in training human judges to perceive lexical diversity just the way the researcher wants them to rather than the way they may be naturally inclined to do. In other words, it could result in an artificially crafted construct.

Regarding the multiple-regression analysis, it should be noted that the only index with a significant part and partial correlation is rarity (Beta = 0.238, t = 2.200, Partial = 0.318, Part = 0.227). The other indices suffer from collinearity, with volume being the main problem (VIF = 36.05). When the same multiple regression analysis was run using a stepwise procedure, the analysis constructed a model of lexical diversity consisting of only volume and rarity. As it turned out, the two-property model had a very slightly higher adjusted R^2 (R^2 = 0.49 (F[1,47] = 24.28, p<0.001) than the full model. Clearly, though, volume and rarity are not the only useful predictors of lexical diversity; when I ran a third multiple-regression analysis with five of the six properties – excluding only volume from the model – there was no longer any collinearity problem, and the result was an adjusted R^2 value only slightly lower than before (R^2 = 0.47, F[5,44] = 9.78, p<0.001). The results of the three regression analyses thus seem to suggest that all six properties are indeed useful for predicting human judgments of lexical diversity, but also that the current operationalization of these properties is in need of further refinement in order to avoid problems of (multi)collinearity. More simply stated, the present results offer support for the proposed six-dimensional construct of lexical diversity, but they also suggest the need to devise purer, more robust measures of some or all of the six properties.

6. Conclusions

In this paper, I have suggested that phenomena in the subjective domain of lexical deployment are complex, multifaceted, and closely associated with the subjective construct of language proficiency. Lexical redundancy is one of these phenomena, and lexical diversity is proposed here as its positive counterpart, although I do not suggest that lexical diversity is necessarily the exact opposite of redundancy in the

way that variability is (at least in some measures) the mirror image of repetition. As I have attempted to characterize it, lexical diversity exhibits a great deal of multidimensional complexity, and this complexity is not sufficiently captured by measures that reduce texts to first occurrences and repetitions, or even by measures that take into consideration the full frequency spectrum of a text. Future measures of lexical diversity need to account for all relevant dimensions of complexity bound up in lexical diversity, particularly the six properties I have proposed in this paper. The model I have proposed characterizes lexical diversity as a perceptual phenomenon that, like color and redundancy, is subjective as a holistic phenomenon, but which nevertheless has measurable component properties that can be calibrated with human perception in a way that allows it to be predicted on the basis of objective measurements.

Although the empirical results presented in this paper are only preliminary and suffer from problems associated with short texts, non-optimized measures, and moderate levels of inter-rater reliability, the evidence does suggest that all six of the proposed properties contribute in some way to the perception of lexical diversity and the related subjective construct of language proficiency. Clearly, a good deal more work will be needed to sort out exactly how these variables interrelate and how they can best be measured by themselves and in concert with one another. In fact, the construct will not be fully defined until we can specify more precisely how these dimensions interact with one another, how they can be made orthogonal to one another, whether they contribute equally to the perception of lexical diversity, and whether the perception of lexical diversity varies across contexts and raters.

I anticipate that some researchers will reject the proposed construct. Perhaps the strongest argument against the model would be the claim that human judgments are not the standard against which to measure the usefulness of existing lexical indices. A loosely related argument is that terminology is arbitrary, and that there is no need to restrict the range of indices or types of indices that the term *lexical diversity* can be applied to. As a rebuttal to the first argument, I would ask the following: If a consensus of human judgments is not the ultimate authority on what lexical diversity is, then who or what is? I would also return to the distinction between repetition and redundancy, for which there is no controversy that the latter construct is (a) grounded in perception and (b) of more theoretical and practical value than the former. I would likewise point to the related construct of language proficiency, which is also grounded in perception, and whose objective measures are ultimately validated by how well they conform to human judgments. In short, when it comes to language phenomena, human perception does matter and should usually be given priority, as I believe it should in this case.

An additional part of my rebuttal to the first argument would be a return to what I said earlier about how most of the existing lexical measures of what I call lexical variability actually are already influenced by multiple factors, including especially volume and evenness. Even if other researchers were to reject the notion that lexical diversity is grounded in perception, there is still good reason to create a more complex model of the phenomenon built on a well-developed understanding of how all of these different dimensions of complexity interact with one another in the objective domain.

In reaction to the argument that terminology is arbitrary, I would point out that one of the most compelling reasons for using the term *lexical diversity* in the way that I propose is that this phenomenon manifests many of the same characteristics as other forms diversity. In fact, I would argue that it shares the same six properties with all other forms of diversity (e.g., color diversity, ecological diversity, racial diversity) (see Chao & Jost, in press, regarding the properties of biodiversity).

An extension of the argument related to terminology is that my proposal confuses lexical diversity with lexical richness, the latter of which already encompasses both variability and rarity and is broad enough to subsume all of the other proposed properties, as well. However, this argument is not particularly compelling, for the following reason. To begin, the term *lexical richness* has suffered a great deal of terminological drift since it was first used by Yule (1944) to refer to the number of words an author has in his or her mental lexicon. Later, it was used interchangeably with lexical diversity (e.g., Honoré, 1979; Vermeer, 2000), and now it is used loosely as a cover term that subsumes lexical diversity along with any other lexical indices a researcher might be interested in (e.g., Malvern et al., 2004; Read, 2000). Although I agree with Yu's (2010) concerns about nomenclature confusion, I believe the cause of this problem is that these terms have not been anchored to clearly articulated construct definitions. This paper proposes a construct definition of lexical diversity that may seem similar in scope to the current meaning of lexical richness, but it nevertheless differs in the sense that the former is now very clearly defined within set parameters whereas the latter is not. I think our field would do well to return lexical richness to something very similar to its original meaning, and to use both *richness* and *diversity* in a way that is compatible with other fields. In ecology, for example, richness refers narrowly to the number of species (types), whereas diversity is a larger construct that subsumes richness along with several other dimensions of diversity (see, e.g., Chao & Jost, in press; Jarvis, 2013). This is how I recommend that these terms be used in our field, as well.

In the future, it will be necessary to compile and analyze a much larger database of texts that have been rated for lexical diversity by human judges. One problem I have already alluded to is the challenge of reaching satisfactory levels of consistency (both intra-rater and inter-rater reliability) in the perception of lexical

diversity. The training and norming of raters has certain disadvantages, as I have mentioned, so it might be useful to investigate whether both intra- and inter-rater reliability improve as the result of rating experience alone. If it does, then this will provide valuable support to the notion that lexical diversity is a real, perception-based phenomenon that is subject to general principles that apply across individuals. Improved rater reliability will also of course improve the precision and confidence levels of the statistical tests that are run on the data, and thus facilitate the calibration of the adopted measures with human judgments.

Regarding the measures themselves, future work will benefit from anchoring at least some of these – particularly rarity, dispersion, and disparity – in some of the same contextual constraints that have been found to govern redundancy (e.g., Bazzanella, 2011; Reynolds, 1995). Although I do not believe that lexical diversity is simply the opposite of redundancy, it may be that the opposite of redundancy is nevertheless fully subsumed within lexical diversity. If so, then it will be necessary to incorporate all that is known about redundancy into the construct definition of diversity. The ultimate frontier for lexical diversity research will be an accounting of how each word, situated in its various levels of context, contributes to or detracts from the perceived lexical diversity of the text as a whole.

Acknowledgements

I express my sincere thanks to Max Rhinehart, Elizabeth Story, and Nataliya Telegina for preparing the texts and collecting the human judgments of lexical diversity.

References

Arnaud, P.J.L., & Savignon, S.J. (1997). Rare words, complex lexical units and the advanced learner. In J. Coady & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 157–173). Cambridge, UK: CUP.

Athanasopoulos, P. (2009). Cognitive representation of colour in bilinguals: The case of Greek blues. *Bilingualism: Language and Cognition, 12*, 83–95.

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford, UK: OUP.

Bachman, L.F., & Palmer, A.S. (1996). *Language testing in practice*. Oxford, UK: OUP.

Barraclough, T.G., Hogan, J.E., & Vogler, A.P. (1999). Testing whether ecological factors promote cladogenesis in a group of tiger beetles (Coleoptera: Cicindelidae). *Proceedings of the Royal Society of London B, 266*, 1061–1068.

Bazzanella, C. (2011). Redundancy, repetition, and intensity in discourse. *Language Sciences, 33*(2), 243–254.

Carroll, J.B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *The Psychological Record*, *2*(16), 379–386.

Carroll, J.B. (1964). *Language and thought*. Englewood Cliffs, NJ: Prentice Hall.

Chao, A., & Jost, L. (In press). *Diversity analysis*. London: Taylor & Francis.

Chotlos, J.W. (1944). Studies in language behavior: IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs, 56*, 75–111.

Clauset, A., Shalizi, C.R., & Newman, M.E.J. (2009). Power-law distributions in empirical data. *Society for Industrial and Applied Mathematics, 51*(4), 661–703.

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.

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

Crossley, S.A., Salsbury, T., McNamara, D.S., & Jarvis, S. (2011b). What is lexical proficiency? Some answers from computational models of speech data. *TESOL Quarterly, 45*(1), 182–193.

Dugast, D. (1979). *Vocabularie et stylistique: I. Théâtre et dialogue, travaux de linguistique quantitative*. Geneva: Slatkine.

Edwards, R., & Collins, L. (2011). Lexical frequency profiles and Zipf's Law. *Language Learning, 61*(1), 1–30.

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*, 139–155.

Ferrer i Cancho, R., & Gavaldà, R. (2009). The frequency spectrum of finite samples from the intermittent silence process. *Journal of the American Society for Information Science and Technology, 60*(4), 837–843.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Ginebra, J., & Puig, X. (2010). On the measure and the estimation of evenness and diversity. *Computational Statistics & Data Analysis, 54*(9), 2187–2201.

Goldstein, E.B. (2007). *Sensation and perception* (7th ed.). Belmont, CA: Thomson Wadsworth.

Gould, S.J. (1990). *Wonderful life: The Burgess Shale and the nature of history*. New York, NY: Norton.

Gries, S.T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge.

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Herdan, G. (1960). *Type-token mathematics: A textbook of mathematical linguistics*. The Hague: Mouton.

Honoré, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, *7*(2), 172–177.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57–84.

Jarvis, S. (2012). Lexical challenges in the intersection of applied linguistics and ANL. In Philip M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

Jarvis, S. (2012). Lexical challenges in the intersection of applied linguistics and ANLP. In C. Boonthum-Denecke, P.M. McCarthy, & T. Lamkin (Eds.), *Cross-disciplinary advances in applied natural language processing: Issues and approaches* (pp. 50–72). Hershey, PA: IGI Global.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning, 63* (supplement 1), 87–106.

Johnson, W. (1939). *Language and speech hygiene: An application of general semantics*. Ann Arbor, MI: Edwards Brothers.

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

Landauer, T.K., McNamara, D.S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.

Laufer, B., & Nation, I.S.P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307–322.

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

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

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. New York, NY: Palgrave MacMillan.

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

McCarthy, P.M. (2005). *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [Microfiche]. Doctoral dissertation, University of Memphis.

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.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*, 323–338.

Meara, P. (2009). *Connected words: Word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect*, *16*, 5–19.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam: John Benjamins.

Michéa, R. (1971). De la relation entre le nombre des mots d'une fréquence déterminée et celui des mots différents employés dans le texte. *Cahiers de Lexicologie*, *18*, 65–78.

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

O'Grady, W., Archibald, J., Aronoff, M., & Rees-Miller, J. (2010). *Contemporary linguistics: An introduction* (6th ed.). Boston, MA: Bedford/St. Martin's.

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

Reynolds, D.W. (1995). Repetition in nonnative speaker writing: More than quantity. *Studies in Second Language Acquisition*, 17(2), 185–209.

Sichel, H.S. (1975). On a distribution law for word frequencies. *Journal of the American Statistics Association*, *137*, 25–34.

Sichel, H.S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist*, *11*, 45–72.

Simpson, E.H. (1949). Measurement of diversity. *Nature, 163*, 168.

Thomson, G.H., & Thompson, J.R. (1915). Outlines of a method of the quantitative analysis of writing vocabularies. *British Journal of Psychology*, *8*, 52–69.
Tuldava, J. (1993). The statistical structure of a text and its readability. In L. Hrebicek & G. Altmann (Eds.), *Quantitative text analysis* (pp. 215–227). Trier: Wissenschaftlicher Verlag Trier.
Tuldava, J. (1996). The frequency spectrum of text and vocabulary. *Journal of Quantitative Linguistics, 3*(1), 38–50.
Turlik, J. (2008). *A longitudinal study of vocabulary in L2 academic English writing of Arabic first-language students: Development and measurement*. Unpublished PhD dissertation, University of the West of England, Bristol.
Upton, G., & Cook, I. (2006). *Oxford dictionary of statistics* (2nd ed.). Oxford, UK: OUP.
Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65–83.
Walker, S.E. (2011) Density and dispersion. *Nature Education Knowledge 2*(9), 3.
Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics 31*(2), 236–259.
Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge, UK: CUP.
Zipf, G.K. (1935). *The psycho-biology of language*. Boston: Houghton Mifflin.
Zipf, G.K. (1937). Observations of the possible effect of mental age upon the frequency-distribution of words from the viewpoint of dynamic philology. *Journal of Psychology, 4*, 239–244.

Appendix A

Sentences and paragraphs used in Task 1:

1. (Variability manipulated)
   Jane asked for some more food, and Jane got some more food.
   Jane ate biscuits with gravy last night because she was really hungry.
2. (Volume manipulated)
   We run up and down the slope of that hill every morning before sunrise.
   We run every morning.
3. (Evenness manipulated)
   I walked, you walked, Jim walked, and Susan walked, but Jane drove.
   Some of us walked, and some of us came in Jane's car.
4. (Formal disparity manipulated)
   I go there on Fridays, you go there on Saturdays, and Jim will go there tomorrow.
   I go there on Fridays, Jim goes there on Saturdays, and you are going there tomorrow.
5. (Semantic disparity manipulated)
   I would characterize her as honest, truthful, and sincere.
   I would characterize her as smart, energetic, and pretty.

6. (Rarity manipulated)
   The pupil inquired about how to perform the task.
   The student asked about how to do the assignment.

Appendix B

Paragraphs used in Task 2:

(High volume)
There was a girl walking along the street who was alone and hungry. She came to a bakery where bread was being delivered, and she stole some bread from the delivery truck when no one was looking. She tried to run away, but she ran into a man, and they both fell down. That gave the police enough time to find her and catch her.

(Baseline text)
There was a girl who was alone and hungry. She stole some bread from a bakery and tried to run away, but she ran into a man, and they both fell down. That gave the police enough time to find her and catch her.

(High variability)
A lonely, hungry girl stole some bread from one particular bakery. She tried running away, but then bumped into someone and fell on top of him. That gave police officers enough time to find her. Before too long, they did catch the young woman.

(High rarity)
A destitute and lonely young female stole a loaf of bread from a bakery. She attempted to flee, but she collided with a man who was walking toward her, and both of them fell down. In the meantime, a policeman arrived and detained them.

(Low evenness)
One lonely and hungry girl was walking and looking for food. She came to a bakery and stole some bread and started running. She ran into someone and fell down, and that gave the baker and police enough time to find and catch her.

(Low disparity)
There was a girl who was hungry and thirsty. She stole some bread from a bakery because of hunger, and then she ran into a woman, and both females fell down. That gave the police enough time to find her and to catch her.