LANGUAGE LEARNING
A Journal of Research in Language Studies

Language Learning ISSN 0023-8333

Capturing the Diversity in Lexical Diversity

Scott Jarvis

Ohio University

The range, variety, or diversity of words found in learners' language use is believed to reflect the complexity of their vocabulary knowledge as well as the level of their language proficiency. Many indices of lexical diversity have been proposed, most of which involve statistical relationships between types and tokens, and which ultimately reflect the rate of word repetition. These indices have generally been validated in accordance with how well they overcome sample-size effects and/or how well they predict language knowledge or behavior, rather than in accordance with how well they actually measure the construct of lexical diversity. In this article, I review developments that have taken place in lexical diversity research, and also describe obstacles that have prevented it from advancing further. I compare these developments with parallel research on biodiversity in the field of ecology, and show what language researchers can learn from ecology regarding the modeling and measurement of diversity as a multidimensional construct of compositional complexity.

Keywords lexical measures; vocabulary acquisition; ecological approaches; biodiversity; human judgments; Shannon's index; Simpson's index

Introduction

I am fortunate to have a scenic view from my office window. Through the window, I see a vast assortment of beautiful trees both up close and also on the distant hills. The trees represent numerous species: pine, spruce, cedar, fir, juniper, maple, walnut, oak, sycamore, magnolia, and dozens of other species that I have not yet learned to identify. Through the window, I also see a great deal of variety in the manmade structures in view: some made of brick, some made of wood, and some with stucco or vinyl exterior walls. The variety of colors—both in nature and in the man-made structures—is also quite dazzling, and it is impressive how well the colors complement one another.

Correspondence concerning this article should be addressed to Scott Jarvis, 383 Gordy Hall, Department of Linguistics, Ohio University, Athens, OH 45701. E-mail: jarvis@ohio.edu

DOI: 10.1111/j.1467-9922.2012.00739.x

The phenomenon that so bedazzles me as I look through my window can be described as diversity, a condition most dictionaries define in relation to variety and differences. In more precise terms, diversity is a condition or quality that arises from the juxtaposition of multiple varied, or different, elements within the same domain, such as a field of view. However, it is also more than this because diversity is affected by the specific relationships among the elements in the domain, such as how well they complement one another, as well as by the unique contribution to diversity that each element makes. For example, the stunning blue Atlas cedar just outside my window and the blooming magnolia tree in the background clearly add more to the diversity of my view than any of the other trees around them.

The general question I address in this article is whether the definition and observations of diversity just offered have anything whatsoever to do with the notion of lexical diversity—a construct that, like diversity in its general sense, is usually defined in relation to variety and differences (for definitions of lexical diversity, see, e.g., Carroll, 1938; Malvern, Richards, Chipere, & Durán, 2004). This article also deals with the relationship between diversity and richness, two notions that have similar origins but whose meanings have become increasingly differentiated—in different ways—in two fields of research whose paths have intersected a time or two, but which have taken separate routes in the modeling and measurement of both richness and diversity. I will argue that one of those fields chose a better route, a route that has traversed through many other fields and has shown a commitment to gleaning from their discoveries, and a route that illustrates the value of developing a theoretical understanding of diversity in its literal sense rather than using *diversity* merely as a term of convenience. As I will explain, the field that chose the better route is not our own.

Before proceeding further, I should say a few words about how the terms lexical diversity, lexical sophistication, and lexical richness are generally understood by researchers today, and why these notions are important to the fields of first and second language acquisition, bilingualism, multilingualism, and language testing and assessment. The term lexical diversity is used interchangeably with the terms lexical variability, lexical variation, and lexical variety (see Engber, 1995). These terms are usually operationalized into measures designed to capture the proportion of words in a language sample that are not repetitions of words already encountered. Lexical diversity is thus often understood as the inverse of the word repetition rate (cf. Carroll, 1938). Lexical sophistication, in turn, has to do with the use of words that are not among the most frequent in the language (e.g., *ask* vs. *request*; *poor* vs. *destitute*), and which are therefore assumed to reflect more advanced levels of

vocabulary knowledge (e.g., Linnarud, 1986). Finally, the term lexical richness was originally used to refer to the number of words in a person's mental lexicon (Yule, 1944), but has subsequently been used by numerous scholars to refer to the number or variety of words encountered in a language sample (e.g., Daller, Van Hout, & Treffers-Daller, 2003; Tweedie & Baayen, 1998). This meaning of lexical richness is essentially synonymous with the meaning of lexical diversity just given. However, many researchers in the past few years have followed the lead of Engber (1995) and Read (2000) in using the term lexical richness as a superordinate term that covers practically all lexical constructs and their associated measures—including but not limited to lexical diversity and lexical sophistication. In the present article, I point out some of the problems with the existing terminology and with the construct definitions that underlie them (see also Yu, 2010). In the meantime, it is important to point out that research involving lexical measures has produced valuable findings concerning how learners' word choices contribute to the complexity and quality of their language use, and it has also shown that such measures serve as useful indices of learners' levels of language proficiency and stages of acquisition (see, e.g., Malvern et al., 2004).

Although the quality of learners' language performance ultimately depends on the appropriateness of the specific words they choose in specific contexts, researchers who develop and use lexical measures attempt to identify the footprints of such choices at a more abstract level. This approach has already delivered a number of valuable tools and findings (see Jarvis & Daller, in preparation); the purpose of the present article is to suggest ways in which these tools can be refined even further. Because of space constraints and in order to maintain a narrow focus, I do not deal here with the technical details of lexical measures, such as the specific procedures and formulas used to calculate them, whether the basic unit of measurement should be a lexeme or lemma, and whether the basic unit of measurement should include multiword lexical items (e.g., *look at*, *on the contrary*). Discussions of these types of technical issues can be found inter alia in Jarvis and Daller (in preparation) and Malvern et al. (2004).

The present article is structured as follows. First, in the section titled *Route 1: The Path of Quantitative Linguistics,* I briefly describe the route taken by linguists in defining, modeling, and measuring lexical diversity and lexical richness. While acknowledging the profound insights and statistical sophistication our field has shown, I also describe some of its unnecessary obsessions and some of the important landmark discoveries in other fields it has overlooked or prematurely dismissed. Next, in the section titled *Route 2: The Path of Ecology,*

I briefly describe the advances made by ecologists regarding the modeling and measurement of biodiversity and species richness. Crucially, I point out that ecology views diversity as a multidimensional phenomenon whose indices are highly complex but are nevertheless anchored in researchers' intuitions about the nature of diversity. In the final section, titled *Introducing Diversity to Lexical Diversity*, I attempt to show how the principles and properties of biodiversity discovered by ecologists have corollaries in both the objective and subjective domains of lexical diversity. It is relevant to mention that Meara and Olmos Alcoy (2010) have also recently recognized the relevance of ecological models to vocabulary research, though their study focused on the use of a particular ecological sampling method for estimating the number of words a person knows, and did not deal directly with lexical diversity.

Route 1: The Path of Quantitative Linguistics

Current notions of lexical diversity and lexical richness can be traced back to a fairly intensive period of groundbreaking vocabulary research in the fields of linguistics, psychology, and statistical literary analysis that spanned from about 1935 to 1944. As I will describe shortly, studies on lexical diversity and lexical richness have continued in these fields since then, and have also expanded to other related fields, such as first and second language acquisition, bilingualism, and multilingualism, language testing and assessment, and aphasia and other language disorders. In this section, I refer to the areas of inquiry that deal with lexical diversity in all of these fields, collectively, as quantitative linguistics.

What is so important about the period between 1935 and 1944 is that this is when the core foundational research in this area took place, when the relevant terms were coined and defined, and when the agenda was essentially set concerning why researchers need measures of lexical diversity and richness, what they might reveal, and what obstacles need to be overcome in their measurement. In 1935, the American linguist George Zipf observed what has come to be known as Zipf's law, a power law that has since been found to govern a number of phenomena in the physical world beyond linguistics, but which began as the observation that the most frequent word in a text accounts for roughly one-tenth of the entire text, the second most frequent word accounts for roughly one-twentieth, the third most frequent accounts for roughly one-thirtieth, and so forth (Zipf, 1935). The most direct implication of this law is that the rank and frequency of a word are inversely related; however, the principle also has other implications. In 1937, Zipf described its implications for the rate with which

the individual words in a text are repeated. His claim, based on Zipf's law, was that the rate with which a word is repeated is fully predictable on the basis of a constant (i.e., 10) multiplied by the word's frequency rank. For example, the most frequent word in a text can be expected to be repeated at a relatively regular interval of about once every 10 words, the second most frequent at an interval of about every 20 words, and so forth (Zipf, 1937).

Just one year later, the American psychologist John B. Carroll (1938) coined the term *diversity of vocabulary* and acknowledged that the notion underlying this term was essentially equivalent to what Zipf (1937) had earlier called the "average rate of repetitiveness" (see Carroll, 1938, p. 380). Carroll's own definition of diversity of vocabulary stated that it is "the relative amount of repetitiveness or the relative variety in vocabulary" (p. 379). This definition reveals Carroll's belief that lexical diversity is affected by the number of different words (e.g., types) in a text and by how often they are repeated (e.g., the number of tokens per type), but it also reveals his belief that the number of types and their repetition rate can be relativized to the length of the text so that texts of different lengths can be compared on the same diversity scale. This is easier said than done, however; the empirical results of his study showed that neither the repetition rate nor the relationship between ranks and frequencies proposed by Zipf remain constant across differing sample sizes. To help solve the problem, Carroll proposed a diversity equation that attempts to capture the way the repetition rate changes across differing sample sizes. The equation turned out to be inadequate, but Carroll held out hope that a solution would eventually be found.

The studies that followed soon after Carroll (1938) exhibited an awareness of his work and shared his desire to solve the sample-size dependency problem, but most of them did not adopt his use of the term *diversity*. The American psychologist and speech pathologist Wendell Johnson (1939, 1944) proposed the type-token ratio (TTR) as a potentially useful index of what he called vocabulary *flexibility* or *variability*. His solution to the sample-size dependency problem was simply to calculate TTR from a standard number of tokens from each text (e.g., the first 200 words), or alternatively to calculate it as the average TTR over multiple, equally sized subsamples of a text (e.g., the average TTR of as many subsamples of 50 words as can be found in a text). Johnson referred to this latter alternative as the mean segmental type-token ratio (MSTTR). His is one of only two solutions to the sample-size problem that actually do seem to overcome the sample-size dependency problem, though perhaps not fully satisfactorily (see, e.g., Covington & McFall, 2010; Malvern et al., 2004). I will describe the second one shortly.

In the meantime, one of the most influential contributions to the debate about the measurement of the repetition rate came from the British statistician George Udny Yule. After his retirement, Yule developed an interest in the special problems and sources of fallacy faced in the statistical analysis of literary vocabulary. He was aware of Zipf's (1935) work and of the sample-size problem. Given that Yule was a statistician by training and profession, it is perhaps not surprising that his (1944) solution to this problem was immensely more sophisticated than those proposed by Carroll (1938), Johnson (1944), or any of their contemporaries. Yule devoted a whole book to the solution, but it boils down to a single formula involving a sum of probabilities calculated from the number of types occurring at each level of frequency (i.e., the number of words occurring once in a text, twice in a text, and so forth). Although some researchers have found Yule's solution to be uninterpretable (see Malvern & Richards, 1997, for references to such studies), those who understand the mathematics behind it recognize that it represents the probability that, during the random selection of words from a text, the same word will be chosen twice in succession (Baayen, 2001, p. 25). Because a higher probability of choosing the same word twice in succession is essentially equivalent to a higher repetition rate, Yule's Characteristic Constant (i.e., the coefficient in his formula) can be used as an index of the rate of word repetition in a text. In the same work, Yule introduced the term *richness of vocabulary*, whereby he meant "the wealth of words at [the author's] command" and "the total words in their treasure-chests" (p. 83). Comparing this definition with Carroll's (1938) definition of vocabulary diversity, it is interesting to see that the two terms originally had very specific and nonoverlapping meanings. Diversity of vocabulary referred to the observed relationship between types and tokens in actual samples of language use, whereas richness referred to the number of types in a person's mental lexicon.

The studies just described, and many others published between 1935 and 1944, provided the field of quantitative linguistics with the core of its current terminology and also set it on its course of viewing lexical diversity as a matter of word repetition—as a matter of the proportions of types and tokens—and set it in pursuit of developing a measure of the repetition rate that does not vary by text length. As I will attempt to explain next, this area of inquiry is now 60+ years and hundreds of studies down the road, but that road has not taken us very far from where things were in 1944. A narrow focus on solving the text-length problem has prevented language researchers from taking advantage of important advances made in other fields.

Two very important studies were published in the late 1940s in other fields. In 1948, the American mathematician and electrical engineer Claude Shannon published a landmark article that gave rise to a new field: information theory. Among other things, his article introduced the notion of information entropy, which refers to the degree of uncertainty (or unpredictability) in a message. The information entropy of a message can be understood as the amount of information the message contains, and Shannon developed a way to measure this. His insights seem to have gone almost completely unnoticed by the field of quantitative linguistics except for a brief period during the 1950s–1960s when Herdan (1958) noted a mathematical relationship between Yule's and Shannon's indices, and when a small number of other quantitative linguists attempted to apply Shannon's ideas to the quantification of redundancy (see Těšitelová, 1992, pp. 65–66).

The other important study from the late 1940s was a very short article published in 1949 by British statistician Edward H. Simpson. The article was published in the journal *Nature*, and its intended audience appears to have been ecologists. The important points for now are that (a) Simpson showed that Yule's index is an outcome of the binomial distribution, (b) Yule's formula can be substantially simplified, and (c) the simplified index can be used as a measure of the diversity or concentration of categories (e.g., species) within a population. Like Yule's index, Simpson's index reflects the probability of randomly choosing two individuals of the same category (e.g., the same word type or the same species) twice in succession. Unlike Yule's index, however, Simpson's index outputs a precise probability rather than a value located on a scale that is difficult to interpret. A simple transformation of Simpson's index by subtracting it from 1 gives an even more intuitive measure of diversity, which represents the precise probability that any two individuals sampled randomly in succession will belong to different types or species. This index is referred to as the Gini-Simpson index (e.g., Jost, 2006), and it is one of the most frequently used indices of biodiversity in the field of ecology. I will say more about this in the following section. In the meantime, it is important to note that neither Simpson's index nor the Gini-Simpson index appears to have been adopted as a measure of lexical diversity (but see Baayen, 2001), although some new measures of lexical diversity, such as the *vocd D* measure and HD-D are built on similar principles of probability (see McCarthy & Jarvis, 2007, 2010). The problem with the new measures, though, is that they are in many respects simply reinventions of a wheel that was already invented in the 1940s, and they share many of the same shortcomings—shortcomings that are not very well

understood by language researchers, but which have been recognized by the field of ecology, as I will describe in the next section.

The studies that have proposed solutions to the sample-size problem are too numerous to list, but it is important to recognize that there have been a number of researchers who have worked intensively on this problem, including in chronological order Guiraud (1954), Herdan (1964), Carroll (1967), Sichel (1975), Honoré (1979), Sichel (1986), Tweedie and Baayen (1998), Jarvis (2002), Malvern et al. (2004), McCarthy and Jarvis (2007), and many others. According to Baayen (2001), only Yule's and Simpson's indices, and a few multiparameter models "are truly independent of sample size" in theory (p. 211). In practice, however, even these indices turn out not to be independent of sample size. The only measures that do seem to remain relatively constant across wide ranges of sample sizes are Johnson's (1944) MSTTR, mentioned earlier, and McCarthy's (2005) measure of textual lexical diversity (MTLD) (e.g., Covington & McFall, 2010; Malvern et al., 2004; McCarthy & Jarvis, 2010). These two measures are essentially mirror images of each other. MSTTR holds the sample size constant while calculating the mean TTR across different segments of a text, whereas MTLD holds TTR constant (usually at .72) while calculating the average number of words in any segment of text that remains above the TTR cutoff value. An important problem with both measures is that neither evaluates the text as a unified whole. Consider, for example, a text consisting of four paragraphs of equal length, with each paragraph having an equivalently high TTR. Depending on how MSTTR and MTLD segment the text, both indices are likely to show that the text has a high overall TTR value even if the last three paragraphs in the text are exact copies of the first paragraph. Simply stated, these solutions have still not solved the problem that Carroll (1938), Yule (1944) and many others set out to solve several decades ago.

Some things have changed, however. For example, even with imperfect measures, researchers have discovered that indices of lexical diversity serve as useful predictors of other important constructs, such as language proficiency, language complexity, vocabulary knowledge, and lexical proficiency (Berman, Nayditz, & Ravid, 2011; Crossley, Salsbury, McNamara, & Jarvis, 2011; Housen et al., 2011; Muñoz, 2010; Sauro & Smith, 2010; Yu, 2010). Measures of lexical diversity have similarly been found to serve as useful gauges of the effects of, inter alia, aphasia (Crepaldi et al., 2011; MacWhinney, Fromm, Holland, Forbes, & Wright, 2010) and nonnativeness (Gui, 2010; Kormos, 2011). Thus, it is recognized that indices of lexical diversity are useful, even though language researchers have neglected the question of what it is that they are actually measuring. Unfortunately, lexical diversity indices tend to be

validated in accordance with how well they avoid sample-size effects and/or how well they predict other constructs (e.g., proficiency) (e.g., Malvern et al., 2004; Vermeer, 2000) rather than in accordance with how well they measure the construct they are intended to measure (i.e., lexical diversity). In other words, the problem is not that the existing measures fail to predict language proficiency, aphasia, and so forth; the problem is that they lack construct validity because they have not been derived from a well-developed theoretical model of lexical diversity. In short, language researchers do not know how well they measure lexical diversity, which means that we also do not know the precise role of lexical diversity (as a construct independent of existing measures) in language proficiency, aphasia, and so forth.

Another thing that has changed over the years is the meaning of the term *lexical richness*. Recall that for Yule (1944), this term referred to the number of words (types) in a person's mental lexicon. Honoré (1979) and many others, on the other hand, seem to have equated the term with lexical diversity, whereas Engber (1995), Read (2000), and others have treated it as a superordinate category that includes lexical diversity as one of many lexical characteristics of good writing. In other words, the meaning of lexical richness has expanded from one that was very specific and nonoverlapping with that of lexical diversity, to one that is broader than but fully encompasses lexical diversity. Although terminological changes often accompany theoretical advances in a field, such changes can also be the result of terminological drift that occurs when terms are not anchored to well-developed construct definitions. This, in a nutshell, appears to be the most serious and most lingering problem in lexical diversity research.

Although I have not painted a very bright picture of the state of the art in lexical diversity modeling and measurement, one of the bright spots in this area of research—in addition to the promising empirical findings referred to earlier—is the recognition that lexical diversity is a type of linguistic complexity. Lexical diversity indices are often included with other measures of complexity in studies that investigate proficiency as a composite of complexity, accuracy, and fluency (see Mackey & Gass, 2012, pp. 145–146). However, the complexity inherent in lexical diversity should not be understood merely as an effect of the level of complexity of a speaker's internal linguistic system. The lexical diversity of an utterance or written text produces its own effects on the listener or reader, and these effects can also be understood in relation to complexity. According to Dillard and Pfau (2002), when the complexity of a message does not exceed listeners' ability to comprehend the message, "listeners prefer complexity because it is interesting, and lexical diversity should be

preferred because it represents more complex lexical choice" (p. 374; cf. Zipf, 1935, p. 213). A series of studies summarized by Dillard and Pfau confirm this claim, showing that lexical diversity affects listeners' perceptions of a speaker's credibility, competence, likeability, socio-economic status, and communicative effectiveness. Seen through the prism of these results, where complexity in the lexical composition of produced speech is shown to affect a perceiver's experience, the effects of lexical diversity are now starting to sound similar to the effects of a more general type of diversity that I referred to earlier while recounting the wonder I experience when I look out my window. I believe that viewing lexical diversity as a perceptual phenomenon with measurable objective properties that must be calibrated with elements of perception provides an important way forward for future research in this area.

Route 2: The Path of Ecology

Ecologists also view diversity as a matter of compositional complexity, and they likewise ground their understanding of that complexity in relation to its perceivable properties. Although there have not been any studies in ecology that, to my knowledge, have empirically tested human perceptions of biodiversity, ecologists do have a long-standing tradition of validating their indices of biodiversity in relation to how well these indices corroborate what the researchers intuitively know about diversity (Chao & Jost, in press). Ecologists also recognize that the compositional complexity underlying diversity has multiple distinct but interrelated properties, and that a measure of any one of these properties alone may serve as a convenient and useful index of diversity, but it does not constitute a full, direct measure of diversity by itself. Jost (2006) put it this way: "a diversity index itself is not necessarily a 'diversity'. The radius of a sphere is an index of its volume but is not itself the volume, and using the radius in place of the volume in engineering equations will give dangerously misleading results" (p. 363).

Indices of biodiversity go back at least to Fisher, Corbet, and Williams (1943), but the most commonly used indices of biodiversity are two that were discussed in the preceding section: Shannon's (1948) index of informational entropy and the Gini-Simpson index, the latter of which is a simple transformation of the index of diversity proposed by Simpson (1949). Recall that Shannon's index represents the level of uncertainty (or unpredictability) in the results of a sampling process: "When it is calculated using logarithms to the base two, it is the minimum number of yes/no questions required, on the average, to determine the identity of a sampled species" (Jost, 2006, p. 363). The higher

the uncertainty of what a sampled species is likely to be, the higher the level of diversity. As for the Gini-Simpson index, recall that this measure reflects the probability that any two sampled individuals will represent different species. The higher the probability, the higher the diversity.

Unlike quantitative linguists, who have not found much use for these two indices, ecologists have attempted to derive as much utility as possible out of them. Ecologists have recognized, first of all, that both indices are affected by the number of species in a sample as well as by the distribution of individuals across species. To quantitative linguists, these two effects may seem like undesirable intervening variables, but ecologists consider them to be essential properties of diversity. Regarding the effects of the number of species, Chao and Jost (in press) point out that in the hypothetical case of a plague that attacks a community and reduces it from one million to just 100 species, any meaningful index of diversity should show that the compositional complexity of the community has been radically diminished. Concerning the effects of the distribution of individuals across species, Shannon (1948) demonstrated that entropy is maximized when the number of individuals belonging to each species is perfectly even. This mathematical reality, referred to as evenness (Hill, 1973), also has an intuitive component in that a species that comprises a disproportionately small number of individuals is not effectively contributing to the community as a full species. As described by Chao and Jost (in press), "the evenness of a community affects complexity because very rare species have little contact with the majority of the individuals in an ecosystem, and do not contribute much to the variety of interactions in that ecosystem" (pagination not yet available).

Ecologists use the term *richness* to refer to the number of species within a community. This meaning is very similar to what Yule (1944) meant when he introduced the term *richness of vocabulary*. However, what ecologists are primarily interested in is not the actual number of species per se, but rather the effective number of species in a community (MacArthur, 1965). This value depends on both richness and evenness, and can thus be calculated from either Shannon's or Simpson's index, although the two ways of calculating it lead to somewhat differing results. The former index of the effective number of species is referred to as the exponential of Shannon entropy, and the latter is referred to as the inverse Simpson concentration. One of the profound discoveries by ecologists working in the area of diversity analysis is that species richness, the exponential of Shannon entropy, and the inverse Simpson concentration can all be captured by a single formula: They differ only in relation to a single parameter within that formula. When the parameter is set to 0, the output is

species richness; when the parameter is set to a value of 1, the output of the formula is the exponential of Shannon entropy; and when the parameter is 2, the output is the inverse Simpson concentration. The parameter in question determines how much weight is given to species abundance. Species richness is biased toward species having very few members, whereas the inverse Simpson concentration is biased toward dominant species. The exponential of Shannon entropy, on the other hand, gives equal weight to all species in accordance with how abundant they are. It is for this reason that the exponential of Shannon entropy is considered to be the most useful general measure of the effective number of species in a community (see Chao & Jost, in press).

Another important property of diversity recognized by ecologists is referred to as *disparity*. This property involves the degree of taxonomical uniqueness between species in a community and is described by Gould (1990) as follows: "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). The main point is that diversity is not just a matter of statistical frequencies and proportions of a given set of categories, but it also fundamentally involves the qualitative nature of and relationship between those categories. In a similar vein, ecologists recognize that some species represent higher levels of importance than others within an ecosystem. This distinction can be the result of the centrality of a species within a food web, which gives it a relatively high level of influence over other species within the network. A higher level of importance can also be the result of the uniqueness of a species whose loss could not easily be compensated by other species (Lai, Liu, & Jordan, in press). Although the principle of uniqueness in ecology is usually applied to the functional rather than the aesthetic value of a species, this principle may nevertheless account for my earlier observation that the magnolia and blue Atlas cedar trees outside my window seem to contribute more than any other trees to the diversity of my field of view. (Insights related to node centrality in the interactions among words in a language have recently emerged in the field of quantitative linguistics under the guise of Network Theory; see, e.g., Ferrer i Cancho, 2010. The notion of uniqueness, on the other hand, seems similar to the notion of lexical sophistication—or rareness—described earlier.)

Finally, patterns of species density and dispersion have also been investigated in relation to biodiversity. Density is a matter of how tightly clustered individuals are within a given amount of space (i.e., how many organisms per unit of space), whereas dispersion refers to how species are distributed in space vis-à-vis one another (Walker, 2011). Dispersion is at its lowest where organisms cluster by species, and where species thus segregate themselves from one

another, whereas dispersion is at its highest where all species are equally distributed throughout a defined area. (In language, lexical dispersion would be at its highest where tokens of the same type are spread evenly throughout a text without any areas of concentrated repetition.) In ecology, dispersion provides a direct indication of the relationships among species, but indices of both density and dispersion show a statistical relationship with other aspects of biodiversity (e.g., Mormu, Thomaz, Takeda, & Behrend, 2011; Sagar, Raghubanshi, & Singh, 2003).

Introducing Diversity to Lexical Diversity

Although the brevity of the preceding two sections exaggerates, oversimplifies, and neglects a number of important developments and problems experienced in both fields of research, what I have described so far does nevertheless illustrate that the field of ecology has a far more complex understanding of diversity than quantitative linguistics does. Whereas quantitative linguists tend to view diversity as a matter of statistical frequencies involving types and tokens—or, in other words, in terms of the rate of word repetition—ecologists view diversity as a multidimensional phenomenon whose properties include richness (or number of species), evenness (or the proportional distribution of individuals across species), disparity (or the amount of difference between species), importance (which includes both centrality and uniqueness), density and dispersion (which refer to the spatial distribution of individuals and species), and perhaps additional properties of which I am not yet aware. Both fields view diversity as a matter of complexity, but ecologists have gone much further in modeling and developing measures for the different aspects of that complexity. Ecologists have also held to a literal and intuitive understanding of diversity, and this has resulted in a highly developed, intricate picture of what diversity entails. This understanding has also ultimately prevented ecologists from becoming obsessed with the search for an index of diversity that does not vary by sample size. For ecologists, it is obvious that when other diversity-related variables are held constant, such as evenness, disparity, and so forth, a community with a larger number of species will be more diverse than a community with a smaller number of species.

Could an equivalent statement also be true of lexical diversity? That is, all things being equal, is a text containing a larger number of lexical types more diverse than a text containing fewer types? To find out, I administered a survey to 130 participants, 109 of whom were native English speakers (98 undergraduates, 6 graduates, and 5 others) and 21 of whom were nonnative

English speakers (8 undergraduates, 12 graduates, and 1 other). One of the items on the survey asked the participants to judge which of the following two sentences is more lexically diverse, where lexical diversity was defined for them as "the variety of word use that can be found in a person's speech or writing":

(S1) *We run every morning*.
(S2) *We run up and down the slope of that hill every morning before sunrise*.

These two sentences have differing numbers of types, and also differing numbers of tokens, but one thing they have in common is that, within each sentence, the number of types equals the number of tokens. This means that there is no repetition within either sentence, and thus each sentence can be treated as being maximally diverse. Many existing indices of lexical diversity, such as TTR and its derivatives, produce equivalent values for both sentences, treating both as being 100% diverse. The question, though, is whether there really is such a thing as 100% diversity, or whether, instead, the maximum possible diversity increases with the length of the text, as the potential for more types increases. The answer to the latter question appears to be a resounding "yes" given that fully 90% ($n = 117$) of the participants indicated that sentence S2 is more lexically diverse than sentence S1.

Now, one could of course argue that the participants might have been confused by the task, and that what they were really judging was lexical richness rather than lexical diversity. This argument is compelling on one level, but is misleading on another. On the one hand, it probably is true that the participants did not perform their judgments in accordance with Carroll's (1938) definition of lexical diversity as a repetition-related phenomenon; but, it would probably also be misleading to suggest that the participants did not know exactly what they were evaluating, or that what they were evaluating was not diversity, or that they did not have a firm grasp of what diversity means. Their judgments were, after all, perfectly compatible with how diversity is defined in the field of ecology. Rather than claiming that the participants were judging lexical richness instead of lexical diversity, it would probably be more accurate to state that lexical richness and lexical diversity are tautologous when it comes to the contrast between these two sentences.

A second criticism of these results might be that they are completely inconsequential: The argument is that the concept of diversity that guided the participants' judgments simply is not what we are interested in measuring. Instead, language researchers are interested—as they have been since the 1930s—in objectively quantifying rates of repetition in people's speech and writing because they believe that this somehow reflects the size of their mental lexicon

and gives an indication of the state of their language knowledge and skills. My rebuttal begins with the following question: Why should one believe that repetition rates are optimal predictors of these abilities? Even though people who know fewer words are likely to repeat those words more often, this is not always the case. Language is inherently repetitive on both the grammatical and pragmatic levels, and what matters more than the sheer quantity of word repetitions in a person's language use is when, where, how often, and to what extent those repetitions are uselessly redundant (e.g., Bazzanella, 2011). What I am suggesting, in other words, is that lexical diversity should not be seen as the positive counterpart to repetition, but rather as the positive counterpart to redundancy. Given that many of the facets of redundancy are subjective, those who agree with my proposal will recognize that neither redundancy nor lexical diversity can be measured adequately on an objective level without an understanding of how they work on a perceptual level. In the future, lexical diversity indices should not be validated in accordance with their ability to overcome sample-size effects and/or their ability to predict other variables, such as proficiency ratings, but rather should be validated in accordance with their ability to predict the lexical diversity judgments of human raters. As pointed out by an anonymous reviewer, the approach I am advocating here is grounded in a psychophysics of lexical diversity rather than in the tools and conventions of corpus metrics.

The use of human raters does, of course, introduce a few new complications. For example, it raises the question of whether human raters should be trained before rating the lexical diversity of language samples, and, if so, what they should be told about lexical diversity. My own answer to these questions is that it would be a mistake to train raters on how to judge the lexical diversity of texts when the whole purpose of using human judgments—at this stage of research—is to discover whether humans already have a concept of lexical diversity, and, if they do, what its dimensions are, how those dimensions are weighted vis-à-vis one another, and whether the dimensions and weightings of lexical diversity are consistent across raters. Human raters' lexical diversity judgments will likely be affected by numerous interrelated factors, and ecologists' discovery of multiple aspects of diversity provide a useful point of departure for hypothesizing which properties of word use are likely to be involved. As a starting point, language researchers should consider at least the following properties of diversity:

1. size (number of tokens)
2. richness (number of types)
3. effective number of types (e.g., the exponential function applied to Shannon's index; MacArthur, 1965)

4. evenness (e.g., the degree to which tokens are distributed equally across types)
5. disparity (e.g., the proportion of words in a text that are semantically related)
6. importance (e.g., the relative frequency with which the words in a text occur in the language as a whole; cf. lexical sophistication)
7. dispersion (e.g., the average interval between tokens of the same type)

Assuming that appropriate measures can be developed for each of these properties and that they can be made sufficiently orthogonal to one another to avoid problems of collinearity, language researchers might be able to arrive at a full measure of lexical diversity (rather than a repertoire of loosely related indices) by determining appropriate weights for each of these component properties of diversity and by combining them into a single model of lexical diversity that is calibrated optimally with the lexical diversity judgments of reliable human raters. Small steps have recently been taken toward this goal, and the preliminary results look promising, even though there are important challenges yet to be solved (Jarvis, 2012, in preparation).

The project I am currently working on focuses on the lexical diversity judgments of 20 untrained raters who rated the same 50 written essays (narrative film retells produced by native and nonnative speakers of English). Although they were told only that lexical diversity has to do with the variety of words found in a text, the raters' ratings showed high levels of consistency, with a Cronbach's alpha of 0.91 overall, and pairwise inter-rater reliabilities as high as 0.76 (Pearson's correlation). To determine whether they might have been rating the quality (proficiency) of the texts rather than lexical diversity, I later asked them to rate the texts for proficiency. A correlation test run between their lexical diversity ratings and proficiency ratings produced a Pearson r of 0.70, suggesting that the participants were not rating the same construct both times, but that the two constructs are indeed related. The next step in this project is to try to build a multidimensional model of lexical diversity that will predict the human judges' lexical diversity ratings. The model will consist of operationalized measures of the seven properties listed earlier, and perhaps more. Finding the best ways to operationalize these properties is challenging, and my initial attempts have included Shannon's index as a combined measure of both richness (types) and evenness as well as a measure of semantic disparity that estimates the mean number of synonyms in each text based on the WordNet semantic sense index (wordnet.princeton.edu), a measure of rarity (or sophistication or importance) that calculates the mean rarity levels for the words in each text

based on the frequency ranks of the same words in the British National Corpus (www.natcorp.ox.ac.uk), and a measure of dispersion that calculates the mean distance (or number of intervening words) between tokens of the same type. Although I am still refining the measures, the preliminary results of a multiple regression analysis with these measures as independent variables and the human raters' judgments as the dependent variable, show an adjusted R^2 of 0.86. In other words, the multidimensional model—preliminary as it is—accounts for 86% of the variance in the human raters' judgments. This is an encouraging result that suggests that the approach to the modeling and measurement of lexical diversity described in this article may indeed hold a good deal of promise for future research and practice (e.g., teaching and assessment).

Once the model has been fully developed and calibrated with human judgments, the next step will be to see how well the new multidimensional measure of lexical diversity correlates with and predicts other phenomena, such as vocabulary knowledge, language proficiency, writing quality, and so forth. Given the success that many of the existing indices of lexical diversity have already had in predicting various aspects of language knowledge and performance, it seems safe to assume that a full and fully calibrated measure (not just an index) of lexical diversity will be found to be even more useful. Even if this is not the case, there is little doubt that a fully defined and adequately measured construct of lexical diversity modeled in relation to perceived compositional complexity will enhance people's understanding of the intricate interplay between vocabulary knowledge, word use, and language proficiency.

Revised version accepted 17 September 2012

References

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

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

Berman, R., Nayditz, R., & Ravid, D. (2011). Linguistic diagnostics of written texts in two school-age populations. *Written Language and Literacy*, *14*(2), 161–187.

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

Carroll, J. B. (1967). On sampling from a lognormal model of word frequency distribution. In H. Kučera & W. N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406–424). Providence, RI: Brown University Press.

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

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.

Crepaldi, D., Ingignoli, C., Verga, R., Contardi, A., Semenza, C., & Luzzatti, C. (2011). On nouns, verbs, lexemes, and lemmas: Evidence from the spontaneous speech of seven aphasic patients. *Aphasiology*, *25*(1), 71–92.

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

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

Dillard, J. P., & Pfau, M. (2002). *The persuasion handbook: Developments in theory and practice*. Thousand Oaks, CA: Sage.

Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155.

Ferrer i Cancho, R. (2010). Network theory. In P. C. Hogan (Ed.), *The Cambridge encyclopedia of the language sciences* (pp. 555–557). Cambridge, UK: Cambridge University Press.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, *12*(1), 42–58.

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

Gui, L. (2010). A contrastive study of lexical proficiency between L1 and L2 compositions via computerized assessment. *Foreign Language Teaching and Research*, *42*, 445–450.

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. [The statistical characteristics of vocabulary: An essay in metholdogy.] Paris: Presses Universitaires de France.

Herdan, G. (1958). An inequality relation between Yule's characteristic K and Shannon's entropy H. *Kurze Mitteilungen–Brief Reports–Communications brèves*, *IX*, 69–73.

Herdan, G. (1964). *Quantitative linguistics*. London: Buttersworths.

Hill, M. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*, 427–432.

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

Housen, A., Schoonjans, E., Janssens, S., Welcomme, A., Schoonheere, E., & Pierrard, M. (2011). Conceptualizing and measuring the impact of contextual factors in instructed SLA—the role of language prominence. *IRAL—International Review of Applied Linguistics in Language Teaching*, *49*(2), 83–112.

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 ANLP. In C. Boonthum-denecke, P. M. McCarthy, & T. A. Lamkin (Eds.), *Cross-disciplinary advances in applied natural language processing: Issues and approaches* (pp. 50–72). Hershey, PA: IGI Global.

Jarvis, S. (in preparation). Defining and measuring lexical diversity. In S. Jarvis & M. Daller (Eds.), *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins.

Jarvis, S., & Daller, M. (Eds.). (in preparation). *Vocabulary knowledge: Human ratings and automated measures*. Amsterdam: John Benjamins.

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.

Jost, L. (2006). Entropy and diversity. *OIKOS*, *113*(2), 363–375.

Kormos, J. (2011). Task complexity and linguistic and discourse features of narrative writing performance. *Journal of Second Language Writing*, *20*(2), 148–161.

Lai, S.-M., Liu, W.-C., & Jordán, F. (in press). On the centrality and uniqueness of species from the network perspective. *Biology Letters*, *8*.

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish leaners' written English*. Malmö, Sweden: Liber Förlag.

MacArthur, R. (1965). Patterns of species diversity. *Biological Review*, *40*, 510–533.

Mackey, A., & Gass, S. M. (2012). *Research methods in second language acquisition: A practical guide*. Oxford, UK: Wiley-Blackwell.

MacWhinney, B., Fromm, D., Holland, A., Forbes, M., & Wright, H. (2010). Automated analysis of the Cinderella story. *Aphasiology*, *24*, 856–868.

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

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

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. (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.

Meara, P. M., & Olmos Alcoy, J. C. (2010). Words as species: An alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language*, *22*(1), 222–236.

Mormul, R. P., Thomaz, S. M., Takeda, A. M., & Behrend, R. D. (2011). Structural complexity and distance from source habitat determine invertebrate abundance and diversity. *Biotropica*, *43*, 738–745.

Muñoz, C. (2010). Staying abroad with the family: A case study of two siblings' second language development during a year's immersion. *ITL International Journal of Applied Linguistics*, *160*, 24–48.

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

Sagar, R., Raghubanshi, A. S., & Singh, J. S. (2003). Tree species composition, dispersion and diversity along a disturbance gradient in a dry tropical forest region of India. *Forest Ecology and Management*, *186*(1–3), 61–71.

Sauro, S., & Smith, B. (2010). Investigating L2 performance in text chat. *Applied Linguistics (Oxford)*, *31*(4), 554–577.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

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.

Těšitelová, M. (1992). *Quantitative linguistics*. Amsterdam: John Benjamins.

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, *32*, 323–352.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*, 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*, 236–259.

Yule, G. U. (1944). *The statistical study of literary vocabulary*. Cambridge, UK: Cambridge University Press.

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.