Assessment of LDAT as a Grammatical Diversity Assessment Tool

Scott L. Healy, Joseph D. Weintraub, Philip M. McCarthy,
Charles E. Hall, and Danielle S. McNamara

Department of Psychology
Institute for Intelligent Systems
University of Memphis
Memphis. TN 38152

slhealy@memphis.edu, jdwentrb@memphis.edu, pmccarthy@mail.psyc.memphis.edu, charles.hall@memphis.edu,
d.mcnamara@mail.psyc.memphis.edu

Abstract

The purpose of this study is to evaluate the validity of measuring grammatical diversity with a specifically designed Lexical Diversity Assessment Tool (LDAT). A secondary objective is to use LDAT to determine if the level of difficulty assigned to English as a Second Language (ESL) texts corresponds to increases in grammatical, lexical, and temporal diversity. Other methods of lexical diversity assessment, such as type-token ratio (TTR), have been used with varying accuracy in an effort to determine the complexity or level of texts. We analyzed 120 ESL texts independently assigned by their sources to one of four levels (*Beginner*, *Lower-intermediate*, *Upper-intermediate*, and *Advanced*). We demonstrated that LDAT significantly reflected the grammatical diversity within these texts. While the findings conflicted with the prediction that grammatical and lexical diversity would increase with assigned level, we concluded that the implementation of LDAT in text design could provide reliable assessments of grammatical diversity.

Introduction

Grammatical diversity refers to the range and variety of grammatical structures, such as the syntax and types of clauses and phrases, present in a text; however, there appears to be no current automated method for its assessment. Grammatical diversity may provide a valuable index for text complexity because increased knowledge of the target language allows students to understand a wider range of grammatical constructions (Pica, 1984). At present, one available method for measuring text complexity is lexical diversity (also called lexical variation or lexical variety). Lexical diversity measures the range and variety of vocabulary present in a text (McCarthy & Jarvis, 2007). Although lexical diversity is a useful index for a partial representation of a text's complexity and has been used in a wide range of studies including writing performance and clinical linguistics (Carrell & Monroe, 1993; Grela, 2001), it does not assess the grammatical diversity of a text, which might also vary as a function of a text's complexity or difficulty.

Lexical diversity indices have found their way into pedagogy through assessment and text design (DuBay, 2004). Ferris (1994) notes that lexical variety is one factor that influences scores on English as a Second Language (ESL) student writing, with higher ability students using a wider range of words. As students' knowledge of the target language increases, they are capable of comprehending a wider range of vocabulary and grammatical constructions (Pica, 1984; DuBay, 2004). The wider range, in turn, means that students can successfully negotiate texts that are more diverse. Having both lexical diversity and grammatical diversity measurements to analyze texts may be beneficial because texts may differ in their grammatical and lexical diversity levels. Thus, grammatical diversity and lexical diversity indices may successfully distinguish diversity differences either among texts of the same level or among texts across different levels (i.e., comparing a Beginner text and an Advanced text). If these measurements are sufficient, there would be a substantial need for accurate and reliable representations of both grammatical and lexical diversity to avoid the consequences of assigning students materials that are either too simple or too complex. With accurate indices of grammatical and lexical diversity, instructors could assign texts that more precisely address student needs and abilities.

One of the main advantages of incorporating a grammatical diversity index into text design is the influence such an index may have on student learning. Larsen-Freeman (1975) notes that grammatical constructions are acquired because of the frequency of their use rather than their grammatical complexity. Because frequency of use is more important than grammatical complexity, students may not need to be taught grammar explicitly because increased repetition through increased exposure to grammatical structures, such as through reading, may be sufficient. As reading offers students access to a variety of grammar and vocabulary, grammatical and lexical diversity indices offer instructors and material designers a representation of diversity in a text to ensure the desired frequency and diversity of grammatical constructions.

Because there are no established tools to measure grammatical diversity, it is currently only subjectively assessed in material design. On the other hand, the availability of tools has allowed lexical diversity to be used in many studies. Type-token ratio (TTR) is one of the most common used methods for determining lexical diversity; however, because of the way TTR is calculated (unique items divided by total items), TTR has been shown to have many problems. McCarthy and Jarvis (2007) suggest TTR approaches, as well as most other lexical indices have limited accuracy because text length influences their diversity values. The formula for calculating TTR (unique items, or types, divided by total items, or tokens) is problematic when applied to texts of different lengths. The problem occurs because as texts get longer, the number of tokens increases constantly; however, the number of new types steadily falls, meaning that the calculation of TTR makes longer texts appear to be less diverse.

Jarvis (2002) suggests that a better-suited method for measuring lexical diversity may be found in using curve fitting. Curve fitting is achieved by finding the curve of TTR values from randomly sampled words from a text and then determining the best-fitting curve for the values (Jarvis, 2002; Vermeer, 2000). McCarthy and Jarvis (2007), however, found that while curve fitting is an improvement on simple TTR, its lexical diversity measurement is still susceptible to text length. As a part of that work, McCarthy and Jarvis (2007) developed a then unnamed, hypergeometric probability-based tool for analyzing diversity that we now call the Lexical Diversity Assessment Tool (LDAT). LDAT was created to measure lexical diversity to help address the need for more accurate tools. LDAT selects a sample size from a text and determines the *probability* of successfully selecting each type of item being measured. These measurements are combined to produce the diversity of items within the text. As a result, LDAT differs from traditional methods of analysis because the diversity values are calculated from a sample and based on probabilities instead of from the entire text based on ratios. This difference in approach substantially reduces text length confounds that affect traditional diversity indices.

Although LDAT was designed to assess lexical diversity, it is not restricted to measuring lexical items; rather the diversity of any sample of items can be calculated. For instance, the diversity of active/passive voice, count/non-count nouns, and phrase types (noun, verb, preposition, etc.) can just as easily be measured with LDAT, if the data appropriately formatted. For the purpose of our study, we focused on grammatical diversity. Specifically, we assessed LDAT's ability to calculate *clausal* diversity: the range, and variety of finite verbs' tense and aspect within a text.

In a study that provides one of the first computer-based representations of clausal diversity, Duran et al. (2007) developed an initial method for successfully calculating a clausal diversity related measure using Coh-Metrix (Graesser et al., 2004; McNamara et al., in press). Coh-Metrix is a unique computational tool with over 700 indices for measuring language, text, and readability (for additional information, visit cohmetrix.memphis.edu). The closest existing measure for clausal diversity available from Coh-Metrix is mean temporal diversity: the mean repetition of verb tense and aspect in a text. Using a corpus of 150 texts, Duran et al. (2007) tested Coh-Metrix's index for temporal diversity against human raters. The results of Duran and colleagues' study validated Coh-Metrix's ability to assess temporal diversity. In this study, we use the validated Coh-Metrix measure of temporal diversity to assess LDAT's accuracy in calculating clausal diversity.

The primary objective of this study is to assess whether LDAT is a valid means for determining the clausal diversity of ESL texts. A second objective is to determine whether the measurements of clausal, lexical, and temporal diversities are a means for assigning ESL text level by assessing if their values rise with the increase in designer-assigned text level. We hypothesize that there will be significant clausal, lexical, and temporal diversity differences between text levels. We also predict that clausal diversity, lexical diversity, and temporal diversity will increase as text level increases under the assumption that students reading higher-level texts are capable of comprehending greater diversity.

Corpus

A total of 120 ESL texts were retrieved from free ESL websites and from textbooks used by The University of Memphis' Intensive English for Internationals (IEI) program. We organized these texts into four levels as assigned by their source: Beginner (n=30), Lower-Intermediate (n=30), Upper-Intermediate (n=30), and Advanced (n=30). There were four requirements for text inclusion: 1) texts were labeled according to a specific level; 2) texts contained at least 25 clauses; 3) texts represented available ESL reading materials; and 4) texts could not originate from sources such as news papers (which often contain high numbers of one particle clauses) We chose to include an equal number of textbook (n=15) and internet (n=15) texts for each level to maintain an equal representation of each media.

Classification of Clauses

The goal of this analysis was to measure clausal diversity according to tense and aspect. To provide adequate input for LDAT, two raters annotated clauses using specific criteria (see Tables 1 and 2). In this study, a clause is defined as any segment of text containing a subject and a finite verb. Non-finite verbs (i.e., gerunds, infinitives, participles) were ignored because they do not carry markers identifying aspect or tense. Each rater identified finite verbs and annotated them by finding their respective locations on the tables. For example, *walks* was annotated as present simple (a) because its tense is present and its aspect is simple. The sentence *Even though he* ***was going*** *bald, he* ***knew*** *that he* ***needed*** *to cut his hair every two weeks* contains three finite verbs (in bold). These verbs were annotated according to their respective placement within the chart: past progressive (g) – past simple (e) – past simple (e). In addition, Table 2 only serves as a model for annotation because some modals can refer to different times. For example, the sentences *He* ***could sing*** *when he was young* and *I* ***could get*** *married next year* present two unique times using the same modal. These modals would be recorded as past simple (o) and future simple (s) respectively.

Table 1. Aspect and time reference of markings used to classify clauses

Verb Aspect	Present	Past	Future
Simple	(a)	(e)	(i)
Perfect	(b)	(f)	(j)
Progressive	(c)	(g)	(k)
Perfect Progressive	(d)	(h)	(l)

Table 2. Modal plus aspect and time reference of markings used to classify clauses

Modals +	Present	Past	Future
Simple	(m)	(o)	(s)
Progressive	(n)	(p)	
Perfect		(q)	
Perfect Progressive		(r)	

Inter-rater Reliability

Kappa analyses produced an inter-rater reliability of greater than .998 for each text level, with at least 99% of cases rated the same by both raters (see Table 3). The few discrepancies were resolved through discussion.

Table 3. Inter-rater reliability for text level as determined by Cohen's Kappa

Level	Kappa
Beginner	0.99
Lower-intermediate	0.99
Upper-intermediate	1
Advanced	0.99

LDAT

LDAT is a tool originally designed to assess lexical diversity (see Malvern et al. 2004; McCarthy and Jarvis 2007). However, because LDAT functions by assessing the diversity of strings of entered symbols, there is theoretically no limitation as to the diversity it can assess. LDAT utilizes a hypergeometric distribution (a discrete probability distribution), to determine the likelihood of the occurrence of word types (or symbol types) in text. The method is used in a series of sampling from any given finite population. In the case of LDAT, the population is the text (or, more precisely, the tokens in the text, whether words or symbols representing clausal structures). Because LDAT assesses unique types relative to total tokens, the text can be represented as unique grammatical forms, as it is in this study. The LDAT hypergeometric equation can be represented as follows:

$$P(X = x) = h(x; n, M, N) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

where x represents the instances of any given type in the sample and n represents the size of the sample. In this study, we follow the Malvern et al. (2004) model by sampling from 35 to 50 tokens. M represents the number of types in the entire population, and N represents the total population.

The annotations produced from each ESL reading text were transferred into a corresponding text file. These text files served as LDAT's input. LDAT reads the annotations contained within the files and calculates a diversity value for each one.

Results

We conducted an Analysis of Variance (ANOVA) to assess differences of clausal diversity values obtained from LDAT as a function of text level. As predicted, there was a statistically significant effect of text level for the clausal diversity condition, F (3,116) = 4.52, p = .005, η_p^2 = .105 (see Table 4). We conducted a planned-comparison analysis (Bonferroni) to determine significant differences between specific text levels for clausal diversity. As predicted, we found Beginner to be significantly less diverse than Lower-intermediate (p = .039) and Advanced (p = .013).

We conducted an ANOVA to assess differences on lexical diversity values obtained from LDAT as a function of text level. As predicted, there was a statistically significant effect of text level for the lexical diversity condition, F (3,116) = 7.38, p < .001, partial η_p^2 = .160. We conducted a planned-comparison analysis (Bonferroni) to determine significant differences between specific text levels for lexical diversity. As predicted, we found Lower- and Upper-intermediate to be significantly less diverse than Advanced (p = .003). However, contrary to our predictions, we found Beginner to be significantly more diverse than both Lower-intermediate (p = .017) and Upper-intermediate (p = .004). This high diversity may result from the "list-like" nature of beginning texts in contrast to the more cohesive ("repetitive") stories at the

Table 4. Means, F, P, and eta for clausal, lexical, and temporal diversities

Diversity	Beginner	Lower-intermediate	Combined-intermediate	Upper-intermediate	Advanced	F	P	eta
Clausal	4.447 (1.298)	5.673 (1.607)	5.258 (1.715)	4.843 (1.745)	5.839 (2.103)	4.518	< .010	0.110
Lexical	35.860 (.756)	34.968 (1.556)	34.906 (1.333)	34.843 (1.090)	35.889 (.977)	7.381	< .001	0.160
Temporal	0.855 (.088)	0.857 (.083)	0.866 (.071)	0.875 (.058)	0.812 (.098)	3.115	0.029	0.075

Note: Standard Deviation appears in parenthesis

intermediate level, but more research will be needed to verify that possibility.

To provide the data for the second goal of our study, we conducted an ANOVA to assess differences on temporal diversity values obtained from Coh-Metrix as a function of text level. As predicted, there was a statistically significant effect of text level for the temporal diversity condition, *F* (3,116) = 3.11, *p* = .029, partial η_p^2 = .075. We conducted a planned-comparison analysis (Bonferroni Test) to determine significant differences between specific text levels for temporal diversity. Lower values for the temporal diversity index indicate higher diversity. As predicted, we found Upper- intermediate to be significantly less diverse than Advanced (*p* = .024). In addition, we found Beginner to be less diverse than Advanced (*p* = .283) and Lower-intermediate to be less diverse than Advanced *(p* = .213).

As Table 4 shows, there are significant clausal, lexical, and temporal diversity differences between text levels. These differences lead us to hypothesize that clausal, lexical, and temporal diversities comprise three basic characteristics for assessing text level. A correlation analysis between clausal, lexical, and temporal diversities was conducted, revealing that although there was no significant correlation between clausal and lexical diversities, there was a significant correlation between clausal and temporal diversities (see Table 5). The clausal planned-comparison analysis supports LDAT's validity of assessing the clausal diversity of texts because the findings generally follow the prediction that clausal diversity increases with level, with just the one exception: Upper-intermediate clausal diversity contradicted the predicted increase. The tool demonstrated that an increase in clausal diversity is not necessarily consistent with an increase in text level. One possible reason for this result is that the selected texts have been purposely designed for a specific grammatical focus and have a higher frequency of that focus.

Table 5. Correlations between clausal, lexical, and temporal diversities

	Lexical	Temporal
Clausal	0.13	-0.24*
Temporal	-0.06	-

*Correlation is significant at *p* < .01

While both the lexical and temporal diversities of text levels differed significantly, no discernable pattern was found. Therefore, these results suggest that lexical and temporal diversities may not be the primary attribute in determining text level because they were not observed to increase consistently with text level. The lexical and temporal planned-comparison analyses suggests that Beginner and Advanced texts may be more dependent upon both lexical and temporal diversities while Lower- and Upper-intermediate may be less dependent upon lexical and temporal diversities. A possible reason for the lexical diversity occurrence is that Beginner and Advanced texts emphasize vocabulary development more than the Intermediate texts; however, the results of this study suggest that some texts may not be designed with lexical diversity as a crucial component of intended difficulty. The results of the temporal diversity planned-comparison analysis contradicted the assumption that temporal diversity values would correspond to clausal diversity values with increases in text level.

We conducted a *post hoc* ANOVA to assess the differences in clausal diversity values as a function of text level. This ANOVA differed from the previous clausal diversity analysis in that the intermediate levels were combined to form Combined-intermediate, which is represented in Table 4. This analysis was a result of finding the discrepancy in the intermediate levels for clausal diversity. As predicted, there was a statistically significant effect of text level for the clausal diversity condition, *F* (3,116) = 4.913, *p* < .009, partial η_p^2 = .077. As predicted, we found the expected trends and we found the Beginner level to be significantly less diverse than the Advanced level (*p* = .007). Hence, the expected trend seems to emerge; however, there are no clean lines between levels as identified by publishers and writers.

Implications

The main objective of this study was to determine whether LDAT could serve as a useful method for assessing the clausal diversity of texts. The data from this study suggest that LDAT shows promise in indexing clausal diversity. The data further suggest that using LDAT may be an effective method to categorize texts in terms of clausal, lexical, and temporal diversity. Clausal diversity, lexical diversity, and temporal diversity may not provide automatic approaches to assigning levels to non-authentic, pedagogical ESL texts because the results were not entirely consistent with the prediction that an increase in text level would result in an increase in clausal, lexical, and temporal diversities. However, of the three diversity indices, we found clausal diversity to be the most consistent with text level increase. This finding suggests that clausal diversity offers the highest likelihood of success in classifying text level once other factors that influence pedagogical difficulty are identified.

Material designers and publishers are currently limited to methods of determining text level that are influenced by text length. LDAT's independence from such a limitation presents an opportunity for a more accurate and informative assessment of texts. In addition, texts could be developed with the assistance of LDAT to ensure that as students progress through ESL reading levels, they are provided with texts that gradually increase in level of clausal diversity.

Limitations

Although this study produced some significant and potentially important findings, there were limitations. The variety of texts readily available from the Intensive English for Internationals classroom and free online sources could have produced some of the puzzling results, such as the Lower- and Upper-intermediate clausal diversity discrepancy. These results could have been a result of using non-authentic texts that were designed for a specific purpose. Because these texts had to have a high frequency of the desired grammatical clause or lexical item, they may have been purposely designed to be less challenging by being less grammatically or lexically diverse.

Future Work

Lu (2009) developed an automatic analyzer that implements Covington et al.'s (2006) modified D-Level scale to measure the syntactic complexity of sentences. Lu's work utilizes a part-of-speech tagger and parser to analyze sentences. Our study served to demonstrate LDAT's ability to assess a measure of grammatical diversity accurately. Similar to Lu's study, our future work will focus on creating an automated LDAT to remove the time-consuming process required to provide hand coded data. The development of a tool to classify clauses would address this need. If a clause identifier and classifier was to be incorporated into LDAT, the preparation time needed to run analysis would be drastically reduced.. By helping to ensure materials are classified quickly and accurately, this merged software would benefit everyone from professional writers who create ESL materials to students reading those materials.

Although this study was only concerned with active verbs, voice is also an important feature that should be considered when analyzing texts. As such, the passive and active voices could be easily addressed within a similar study, which would have a voice parameter to classify clauses within the text. This study would build on the current study and provide ESL professionals with statistics of the actual distribution of voice across text levels. This feature would be important given that appropriate use of the passive voice is particularly challenging for ESL learners and instructors.

The combination of LDAT and a clause identifier also has implications for assessing student writing. The assessment could benefit by quickly analyzing student writing to determine placement within a school, program, or class. This information could be used much in the same way as speech pathologists assess language proficiency issues.

Acknowledgments

This work was funded in part by the Institute for Education Sciences (R305G020018-02; IES R305A080589). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors also acknowledge the contribution to this project of Scott Jarvis, Scott Crossley, Yuko Fujiwara, and Zhigiang Cai.

References

Carrell, P.L. and Monroe, L.B. 1993. Learning Styles and Composition. *The Modern Language Journal, 77:* 148-162.

Covington, M.A., He, C. and Brown, C. 2006. How Complex is that Sentence? A Proposed Revision of the Rosenberg and Abbeduto D-Level Scale. Research Report 2006-001, CASPR Project, Artificial Intelligence Center, the University of Georgia.

Duran, N.D., McCarthy, P.M., Graesser, A.C., and McNamara, D.S. 2007. Using Coh-Metrix Temporal Indices to Predict Psychological Measures of Time. *Behavior Research Methods,* 29: 212-223.

DuBay, W.H. 2004. The Principles of Readability. Costa Mesa, CA: Impact Information.

Ferris, D.R. 1994. Lexical and Syntactic Features of ESL Writing by Students at Different Levels of L2 Proficiency. *TESOL Quarterly,* 28: 414-420.

Graessar, A.C., McNamara, D.S., Louwerse, M.M., and Cai, Z. 2004. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavioral Research Methods, Instruments, and Computers,* 36: 193-202.

Grela, B.G. 2002. Lexical Verb Diversity in Children with Down Syndrome. *Clinical Linguistics & Phonetics,* 16: 251-263.

Larsen-Freeman, D. 1975. The Acquisition of Grammatical Morphemes by Adult ESL Students. *TESOL Quarterly,* 9: 409-419.

Lu, X. 2009. Automatic Measurement of Syntactic Complexity in Child Language Acquisition. *International Journal of Corpus Linguistics*, 14:3-28.

Jarvis, S. 2002. Short Texts, Best-fitting Curves and New Measures of Lexical Diversity. *Language Testing,* 19: 57-84.

McCarthy, P., and Jarvis, S. 2007. V*ocd*: A Theoretical and Empirical Evaluation. *Language Testing,* 24: 459–488.

McNamara, D.S., Graesser, A.C., McCarthy, P.M., and Cai, Z. in press. *Coh-Metrix: Automated Evaluation of Text and Discourse*. Cambridge University Press.

Pica, T. 1984. L1 Transfer and L2 Complexity as Factors in Syllabus Design. *TESOL Quarterly,* 18: 689-704.

Vermeer, A. 2000. Coming to Grips with Lexical Richness and Spontaneous Speech Data. *Language Testing,* 17:65-83.