2016

Language Sample Length Effects on Various Lexical Diversity Measures: An Analysis of Spanish Language Samples from Children

Morgan Stills
Portland State University

Language Sample Length Effects on Various Lexical Diversity Measures: An Analysis of Spanish Language Samples from Children

By:

Morgan Stills

An undergraduate honors thesis submitted in partial fulfillment of the

requirements for the degree of

Bachelor of Science

in

University Honors

and

Speech and Hearing Sciences

Thesis Adviser

Dr. Maria Kapantzoglou

Portland State University

2016

Introduction

Speech-language pathologists disproportionally misdiagnose children from culturally and linguistically diverse environments as having specific language impairment when in reality language skills are developing typically (Samson & Lesaux, 2009). Lexical diversity is one of the language areas typically assessed to determine the level of language development in children (Owen & Leonard, 2002). Some studies have investigated how well lexical diversity indices differentiate children with and without specific language impairment (Owen & Leonard, 2002; Klee et al, 2004; Wong et al, 2010; Ooi & Wong, 2012). The goal has been to identify a measure of lexical diversity that is effective, yet minutely affected by language sample length. Studies have been conducted in various languages including English (Owen & Leonard, 2002; Heilmann, Nockerts & Miller, 2010; McCarthy & Jarvis, 2010; Kover et al, 2012; Fergadiotis, Wright & Green, 2015), Cantonese (Klee et al, 2004; Wong et al, 2010), Chinese-English bilinguals (Ooi & Wong, 2012) and Spanish-English bilinguals (Simon-Cereijido & Gutierrez-Clellen, 2009; Bedore et al, 2010). Unfortunately, not all of these studies used the same measure of lexical diversity indices or sample demographic. Therefore it is unclear which lexical diversity indices work best when analyzing Spanish language samples. The goal of the current study is to identify the most valid measure of lexical diversity as it pertains to Spanish-speaking children by estimating and comparing five different lexical diversity indices estimated for the same group. Our research will lead to suggestions for accurate assessment of lexical diversity skills in Spanish-speaking children, the larger minority in the U.S. (U.S. Census Bureau, American Community Survey, 2013).

Validity Evidence for Lexical Diversity Measures

As defined by McCarthy and Jarvis (2010), lexical diversity is the range of words in a language sample with a higher number of different words equating to a higher diversity. There are various measures of lexical diversity including the number of different words (NDW), type-token ratio (TTR), HDD, measure of textual lexical diversity (MTLD), moving average type-token ratio (MATTR), and Maas.

One way to determine if a study or analysis provides evidence for the validity of a lexical diversity measure is to see if the target measure is affected by sample length. For example, there was a study performed by Owen and Leonard (2002) that was a comparative analysis of the lexical diversity measures HDD, type-token ratio (TTR) and number of different words (NDW), in order to compare the validity of each measure. The study was conducted with English language samples and found that HDD was a measure of lexical diversity that was less affected by sample size length as opposed to TTR and NDW. In this study, the researchers analyzed language samples from 144 children ranging in age from 2 years 2 months to 7 years 3 months. In another study, Klee et al. (2004) provided strong evidence for HDD as a measure that reflects lexical diversity without the influence of word token size based on Cantonese-speaking children. The authors studied 74 children aged 27-68 months, and the purpose was to examine the relationship between age and the lexical diversity measure HDD. They found that HDD was an improvement on other lexical diversity measures because it wasn't as affected by sample length.

Another type of validity evidence that a measure of lexical diversity indeed measures lexical diversity is that the measure can differentiate between a child with specific language impairment and a child who is typically developing. The reason this is important in a lexical diversity measure, is because children with specific language impairment tend to have smaller vocabularies. An lexical diversity measure that reflects those differences in vocabulary can be

more effective in a clinical setting than a measure that doesn't reflect those differences. Owen and Leonard (2002) found that HDD was sensitive to developmental differences in younger English-speaking children through analyzing language samples from 144 children ages 2;2 to 7;3. In their 2004 study, Klee et al. also assessed whether or not HDD could be an indicator for specific language impairment in Cantonese-speaking children. Their results indicated that HDD is indeed sensitive enough to also serve as an indicator of specific language impairment in Cantonese-speaking children. More recently, Ooi & Wong performed a study in 2012 that analyzed the sensitivity of HDD as an indicator of specific language impairment in Chinese-English bilingual children. This study found that HDD was an effective indicator of specific language impairment, though not the most effective measure analyzed. The results from these three studies provide effective evidence that HDD could be an effective indicator of specific language impairment across languages; thus, HDD could be an accurate measure of lexical diversity in Spanish-speaking children as well.

Comparative Analyses of Lexical Diversity Measures

There have been a number of studies conducted that compare the effectiveness and validity of various lexical diversity measures with one another. All of the following studies were performed on language samples from typically developing English speakers across ages.

Regarding children, Owen and Leonard (2002) performed a study comparing TTR, HDD and NDW, in order to assess which of the three were most susceptible to the length of the language sample. Their analysis included both typically developing children and children with specific language impairment, and they analyzed language samples from 144 children all together, ranging in age from 2;2 to 7;3. Their study found that TTR and NDW were both more affected by language sample length than HDD, which was an important finding and the basis of

further research in the future. In adult populations, McCarthy and Jarvis (2010) performed a study on various printed texts, comparing the lexical diversity measures MTLD, HDD, Maas and TTR. Rather than analyzing speech language samples, these researchers chose to focus on analyzing printed text. Their study was aimed at analyzing the validity of MTLD specifically. They found that it correlates highly with other sophisticated lexical diversity measures, such as HDD, and does not correlate highly with lexical diversity measures that are highly affected by sample length, such as TTR. Based on their analysis, MTLD also remains consistent across sample lengths. Finally in 2015, Fergadiotis, Wright & Green performed a study comparing, MTLD, HDD, Maas, and MATTR to see how accurate the measures were when compared with one another. Their study found that MTLD and MATTR were stronger measures of lexical diversity than HDD and Maas.

Studies where lexical diversity measures were performed on Spanish language samples are few and there are no studies to my knowledge that compare lexical diversity measures in Spanish-speaking children. Also, many of the lexical diversity measures assessed in adults (i.e., MTLD, Mass, and MATTR) have not been tested in language samples of children, although there is evidence that those may be more accurate measures of lexical diversity than NDW, TTR and HDD, which are typically used in children. Based on the results from the studies that performed comparative analyses in English, MTLD, Mass, TTR, MATTR and HDD should be compared as lexical diversity measures using Spanish language samples of children.

Purpose of this Study

This study compared the validity of the lexical diversity measures HDD, MTLD, Maas, TTR and MATTR when analyzing Spanish language samples from typically developing monolingual Spanish-speaking children. Specifically, the current study examined whether there

are differences in lexical diversity estimates when the language sample length changes from 50 words to 100 and 200 words for five lexical diversity indices (i.e., HDD, MTLD, Maas, TTR and MATTR) in 4 to 6 year old monolingual Spanish-speaking children.

Method

Participants

60 monolingual Spanish-speaking children, ages 4 to 6 years, participated in this study. Children were recruited from both public and private schools in three states in Mexico: Mexico City, Querétaro, and Monterrey. The children had not come in contact with indigenous languages. None of the participants had a history of hearing loss, sensorimotor or neurological problems, severe psychological disorders or health problems, according to a parent questionnaire.

Selection criteria. The children were selected based on the following criteria: a) parent report indicated no concern of speech or language impairment; b) the number of grammatical errors per terminable unit (T-unit) in the language sample was below 20% (Restrepo, 1998), and c) 4-year-old children scored above the cut score of 50 on the Bilingual English-Spanish Language Test (BESA; Peña, Gutiérrez-Clellen & Iglesias, 2006), and 5- and 6-year-old children scored within one standard deviation (SD) from the mean or above on two grammatical subtests of the Clinical Evaluations of Language Fundamentals-Fourth Edition, Spanish (CELF-4; Word Structure and Sentence Repetition). The CELF-4 Spanish (Wiig, Semel, & Secord, 2006) was used for children 5- and 6-years old because it has higher sensitivity than BESA for this age range.

Measures

Identification measures. The combination of parent concern, performance on a standardized norm-reference tests and percentage of grammatical errors in the story retell can provide enough information for detecting a potential language impairment. This information had 100% agreement in the clinical judgment of the speech-language pathologists.

Parent Questionnaires. Parent report was used to measure to profile the participant's language use and proficiency, the child's education history, their general health concerning motor, neurological, and psychological evolution, audition issues, maternal and paternal education, and any other speech and language concerns the parents might have about their child.

BESA. The BESA is as standardized norm-referenced test designed as a diagnostic tool for children with potential language impairment who speak Spanish. The morphosyntax subtest was used because it is considered to be accurate between the ages of 4 years and 5 years 11 months in mono- and bilingual children. According to the technical manual, for Spanish-English speaking children between 4 years and 5 years 11 months, the sensitivity of the morphosyntactic subtest is 87.5% and the specificity 100%.

CELF-4 Spanish. Children age 6 were evaluated for potential language impairment using CELF-4 Spanish due to it having higher sensitivity and specificity for Spanish-speaking participants above age 5. The test manual reports sensitivity of 96% and specificity of 87% for the core language score at 1 SD below the mean for Spanish-English speakers. For the monolingual population in this sample, the cut score had to be adjusted one standard deviation considering the clinical report of the two speech-language pathologists, the language samples and previous reports on this measure (Morgan et al. 2009; 2013).

Language samples analyses. A language sample in the form of a story retell was collected from each child to assess language abilities based on the number of grammatical errors

in the language sample (Restrepo, 1998). The clinician read the script of two different books, and then asked the child to retell the story to the tester. Narratives were transcribed and coded by three graduate assistants who used the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2010) computer program to find grammatical errors in each language sample. Semantic, phonological or cohesive errors were not counted as grammatical errors. Instead, omissions, substitutions, additions and word order mixes were considered errors. Any discrepancies were resolved by consensus with a third transcriber. Transcribers are linguists who separated utterances into T-units, and coded for grammaticality per T-unit for indicating potential language impairment status. Grammaticality above 20% errors per T –unit indicated potential language impairment, and those children were excluded from the current study.

Lexical diversity measures. The lexical diversity measures we tested for validity in the Spanish language are HDD, Maas, MTLD, TTR and MATTR. Measure HDD captures how fast the TTR decreases in the sample (Owen & Leonard, 2002). The Maas Index is based on the assumption that the TTR curve can be fitted by a logarithmic curve. MTLD is a measure that analyzes how many consecutive words a certain TTR is maintained. Finally, MATTR calculates the lexical diversity using a moving block that estimates TTRs for each following block of fixed length (Fergadiotis, Wright & Green, 2015).

Analysis

To determine whether length was associated with statistically significant differences in lexical diversity values, we analyzed scores from each index using a repeated measure ANOVA with three levels (50 Words, 100 Words, and 200 Words).

Results

Table 1, Table 2 and Table 3 provide basic statistical analysis by language sample length.

TTR

There was a statistically significant effect of the length, *Wilks' Lambda* = 0.52, F (2, 58) = 53.23, *partial* η^2 = *.948* $p < .001$. The significant main effect was followed up by pairwise comparisons using paired sample *t*-tests. There were statistically significant differences between all length conditions with a large effect size. The mean TTR based on 50 words was higher compared to both TTR means based on 100 and 200 word samples, $t(59) = 20.80$, $p < .001$, and $t(59) = 32.51$, $p < .001$, respectively. In addition, the difference between TTR means based on 100 and 200 word samples was also statistically significant, with the mean TTR of the 100-word sample being higher, $t(59) = 19.97$, $p < .001$. In order to better compare the samples to one another, we had to calculate the effect size or Cohen's d. We calculated this using the *t*-score and dividing that by the square root of our sample size, 60. For TTR comparing the means at 50 and 100 words d=2.69, and when comparing the means of 50 and 200 words d=4.20. For the difference between 100 and 200 word sample lengths, d=2.58.

Maas

Maas scores were also associated with a statistically significant effect as a function of length even though the effect size was substantially smaller compared to TTR, *Wilks' Lambda* = 0.884, F (2, 58) = 3.749, *partial* η^2 = *.116*, $p = .03$. The significant main effect was followed up by pairwise comparisons using paired sample *t*-tests. The mean Maas estimate based on 50 words was lower compared to the mean based on 100 word samples, $t(59) = 2.34$, $p = .02$. The remaining comparisons were not statistically significant after controlling for Type I error. Cohen's d was also calculated for comparisons with Maas. The comparison that was significant was that between the 50 and 100 word samples, and for this d = 0.30.

MTLD

MTLD scores, similarly to Maas scores, were associated with a statistically significant effect as a function of length but the effect size was again very small, *Wilks' Lambda* = 0.874, *F* (2, 58) = 4.191, *partial* η^2 = *.126, p* = .02. The significant main effect was followed up by pairwise comparisons using paired sample *t*-tests. Only the comparison of the 100 and 200 word samples yielded a statistically significant difference, after controlling for Type I error, *t* = 2.66, *p* = .01, d=0.34. The average of the 200-word sample was higher than the average of the 100-word sample.

MATTR

With respect to MATTR, we found a statistically significant effect, *Wilks' Lambda* = 0.563, *F* (2, 58) = 19.39, *partial* η^2 = *.437, p* < .001. The significant effect was followed up with pairwise comparisons. Specifically, the mean MATTR based on 100 words was lower compared to both MATTR means based on 50 and 200 word samples, *t*(59) = 4.08, *p* < .001, d=0.53, and *t*(59) = 5.864, *p* < .001, d=0.76, respectively. The difference between MATTR means based on 50 and 200 word samples was not statistically significant.

HDD

Finally, HDD scores were also associated with a statistically significant effect as a function of length but once again the effect size was small, *Wilks' Lambda* = 0.836, *F* (2, 58) = 5.701, *partial* η^2 = *.164, p* = .005. The significant main effect was follow up by pairwise comparisons using paired sample *t*-tests. Only the comparison of the 100 and 200 word samples yielded a statistically significant difference after controlling for Type I error, *t* = 3.41, *p* = .01. Based on the *t*-score of 3.41 for the difference between the 100 and 200 word samples, Cohen's d=0.44.

Discussion

TTR

As it was expected, the length of the sample significantly affected the TTR, as it was demonstrated by the decrease in the mean TTR as the sample length increased. Large effect sizes indicated that TTR is highly affected by sample length, and therefore TTR is a weak measure of lexical diversity. This result is consistent with the studies by Owen & Leonard (2002), McCarthy and Jarvis (2010), and Fergadiotis, Wright, and Green (2015). Specifically, Owen & Leonard (2002) found that TTR decreased as sample length increased in 2 to 7 year-old English speaking children, McCarthy and Jarvis (2010) found similar results in their analysis of printed texts, and Fergadiotis, Wright, and Green (2015) in their analysis of adult discourse samples. The consistency of the results across language samples from different populations strengthens these findings.

Maas

The length of the language sample also affected Maas, but the effect size was not as large for Mass as it was for TTR. These results indicate that Maas is less affected by language sample length than TTR. The greater difference in lexical diversity estimates occurred between the language samples of 50 and 100 words. These findings suggest that Maas is a stronger measure of lexical diversity because it is significantly less affected by length than TTR. Results are consistent with the study by McCarthy and Jarvis (2010), which indicated the strength of the measure when comparing MTLD, HDD, Maas and TTR in printed texts. The fact that similar results came from studies that analyzed two different types of language samples – English texts (McCarthy and Jarvis, 2010) and Spanish language samples elicited from children – supports the current findings about the validity of Maas.

MTLD

MTLD yielded similar results to Maas, showing a small effect size for being affected by sample length. For MTLD, the statistically significant difference occurred between the 100 and 200 word sample lengths. Our results are consistent with both the findings from McCarthy and Jarvis (2010) and Fergadiotis, Wright and Green (2015) who studied printed text and adult discourse samples respectively. Both studies found MTLD to be a strong measure, minimally affected by sample length.

MATTR

Based on the results of the current study, MATTR also was affected by length of the sample, but not as much as the TTR. The effect size for MATTR was much lower than for TTR. Also, the MATTR dropped for the 100-word condition. According to our data, MATTR isn't the strongest measure in regard to how much influenced it is by language sample length, which is inconsistent with results by Fergadiotis, Wright and Green (2015). Their finding suggested that MATTR was a stronger measure than HDD and Maas.

HDD

HDD was also slightly affected by length. The results indicate that it is more affected by length than Maas and MTLD, and less affected than TTR and MATTR. The only comparison that was statistically significant was between the 100 and 200 word conditions. These results are consistent with the findings of Owen and Leonard (2002), McCarthy and Jarvis (2010), and Fergadiotis, Wright and Green (2015). Our results of HDD being a more effective measure of lexical diversity than TTR because it is less affected by length are consistent with Owen and Leonard (2002). Our results are also consistent with Fergadiotis, Wright and Green (2015) and their finding that HDD is less effective of a measure than MTLD.

To summarize, according to our data, the most effective measures for measuring lexical

diversity in the Spanish language are Maas and MTLD, followed by HDD. The least effective measures are TTR and MATTR. The results of the current study are to a large extent consistent with previous studies on lexical diversity conducted previously on printed text and discourse samples from populations with different characteristics.

In the current study, for the 100-word condition, there was a drop in lexical diversity across all measures. This could be due to the differences in methodology in our study versus other studies that have been performed comparing lexical diversity measures. For our study, we always started at the beginning of the language sample, whereas other studies used the middle of language samples to analyze. This difference could be accentuated because of the way stories are told. The beginning of the story is where there's a lot more description and defining of new terms, then in the middle the storyteller is using a lot of those words they previously defined, and at the end there's more new words being used.

Future Implications

This research could be used as a basis for creating hypotheses in future studies that wish to analyze the effectiveness of lexical diversity measures in languages other than English and Spanish. Also, results from the current study support the use of Maas and MTLD, as the most effective lexical diversity measures in English and Spanish to assess language development in Spanish-English bilinguals and potentially identify language impairment in the target population.

References

Bedore, L. M., Peña, E. D., Gillam, R. B., & Ho, T.-H. (2010). Language sample measures and language ability in Spanish-English bilingual kindergarteners. *Journal of Communication Disorders*, *43*(6), 498–510. http://doi.org/10.1016/j.jcomdis.2010.05.002

Fergadiotis, G., Wright, H. H., & Green, S. B. (2015). Psychometric Evaluation of Lexical Diversity Indices: Assessing Length Effects. *Journal of Speech Language and Hearing Research*, *58*(3), 840. http://doi.org/10.1044/2015_JSLHR-L-14-0280

Heilmann, J., Nockerts, A., & Miller, J. F. (2010). Language Sampling: Does the Length of the Transcript Matter? *Language, Speech, and Hearing Services in Schools*, *41*(4), 393–404.

Klee, T., Stokes, S. F., Wong, A. M.-Y., Fletcher, P., & Gavin, W. J. (2004). Utterance length and lexical diversity in Cantonese-speaking children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *47*(6), 1396–1410.

Kover, S. T., McDuffie, A., Abbeduto, L., & Brown, W. T. (2012). Effects of Sampling Context on Spontaneous Expressive Language in Males With Fragile X Syndrome or Down Syndrome. *Journal of Speech, Language, and Hearing Research*, *55*(4), 1022–1038.

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. http://doi.org/10.3758/BRM.42.2.381

Miller, J. F., Iglesias, A. (2010). Systematic analysis of language transcripts (Version 16) [software]. Available from http://saltsoftware.com/

Morgan, V. L., Mishra, A., Newton, A. T., Gore, J. C., & Ding, Z. (2009). Integrating Functional and Diffusion Magnetic Resonance Imaging for Analysis of Structure-Function Relationship in the Human Language Network. PLOS ONE, 4(8), e6660. http://doi.org/10.1371/journal.pone.0006660

Ooi, C. C.-W., & Wong, A. M.-Y. (2012). Assessing bilingual Chinese-English young children in Malaysia using language sample measures. *International Journal of Speech-Language Pathology*, *14*(6), 499–508.

Owen, A. J., & Leonard, L. B. (2002). Lexical Diversity in the Spontaneous Speech of Children With Specific Language Impairment Application of D. *Journal of Speech, Language, and Hearing Research*, *45*(5), 927–937.

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A. (2006) *Bilingual English Spanish assessment*. San Diego, CA: AR-Clinical Publications.

Restrepo, M. A. (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research*, 41(6), 1398-1411.

Samson, J. F., & Lesaux. (2009). Language-minority learners in special education: rates and predictors of identification for services. *Journal of Learning Disabilities, 42*(2), 148–162.

Simon-Cereijido, G., & Gutierrez-Clellen, V. F. (2009). A Cross-Linguistic and Bilingual Evaluation of the Interdependence between Lexical and Grammatical Domains. *Applied Psycholinguistics*, *30*(2), 315–337.

U.S. Census Bureau, American Community Survey. (2013). Retrieved from http://www.census.gov/acs/www/data/data-tables-and-tools/

Wiig, E. H., Semel, E., Secord, W. A. (2006). *Clinical evaluation of language fundamentals- fourth edition Spanish*. San Antonio, TX: Pearson.

Wong, A. M.-Y., Klee, T., Stokes, S. F., Fletcher, P., & Leonard, L. B. (2010). Differentiating Cantonese-Speaking Preschool Children With and Without specific language impairment Using MLU and Lexical Diversity (D). *Journal of Speech, Language, and Hearing Research, 53*(3), 794–799.

Table 1

Language Sample Length 50 Words

Statistical Analysis	TTR	MATTR	Maas	HDD	MTLD
Average	0.55	0.48	0.30	24.61	20.29
Standard Deviation	0.07	0.20	0.03	2.78	6.78
Minimum	0.38	0.00	0.24	17.59	10.52
Maximum	0.68	0.69	0.38	29.88	40.55

Language Sample Length 50 Words. Statistical analysis conducted for Type-Token Ratio (TTR), Moving Average Type-Token Ratio (MATTR), Maas, HDD, and Measure of Textual Lexical Diversity (MTLD).

Table 2

Language Sample Length 100 Words

Statistical Analysis	TTR	MATTR	Maas	HDD	MTLD
Average	0.43	0.53	0.32	24.55	19.54
Standard Deviation	0.05	0.06	0.08	2.16	5.07
Minimum	0.31	0.41	0.25	19.17	11.42
Maximum	0.56	0.68	0.89	30.25	39.21

Language Sample Length 100 Words. Statistical analysis conducted for Type-Token Ratio (TTR), Moving Average Type-Token Ratio (MATTR), Maas, HDD, and Measure of Textual Lexical Diversity (MTLD).

Table 3

Language Sample Length 200 Words

Statistical Analysis	TTR	MATTR	Maas	HDD	MTLD
Average	0.33	0.56	0.30	25.12	20.50
Standard Deviation	0.04	0.05	0.02	1.88	5.23
Minimum	0.25	0.45	0.26	20.78	13.84
Maximum	0.43	0.71	0.34	30.68	45.17

Language Sample Length 200 Words. Statistical analysis conducted for Type-Token Ratio (TTR), Moving Average Type-Token Ratio (MATTR), Maas, HDD, and Measure of Textual Lexical Diversity (MTLD).