Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Marginalised Children's Vocabulary Index (MCVI) : A PCA-Corpus Approach

Muhamad Fadzllah Zaini

mfadzllah@fbk.upsi.edu.my

Sultan Idris Education University

Article

Keywords: Lexical Diversity, Marginalised Children's Vocabulary Index (MCVI), Corpus Analysis, Principal Component Analysis, Marginalised children

Posted Date: March 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8986740/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Marginalised Children's Vocabulary Index (MCVI) : A PCA-Corpus Approach

Abstract

This study aimed to develop the Marginalised Children's Vocabulary Index (MCVI) as a composite measure of lexical diversity among marginalised children. The study data consisted of 670 school holiday essays covering three levels of education, namely primary school, vocational college and pre-university. To overcome the weaknesses of traditional text length-sensitive measurements, this study integrated the Moving Average Type–Token Ratio (MATTR50) and the Measure of Textual Lexical Diversity (MTLD) using Principal Component Analysis (PCA). The findings show that MCVI has a stable statistical distribution and is relatively independent of the effect of text length. Further analysis found that internal heterogeneity within each level of education was more dominant than the average difference between levels of education. Overall, MCVI has proven to be effective as an individualised linguistic profiling tool and provides an empirical basis for language development analysis as well as more targeted educational intervention planning in marginalised child populations.

Keywords: Lexical Diversity; Marginalised Children's Vocabulary Index (MCVI); Corpus Analysis; Principal Component Analysis; Marginalised children

Pengenalan

"Do you know what the Petronas twin towers in Kuala Lumpur look like?", asked a friend who has just moved from the city to the countryside, to which a friend from the countryside replied, *"Let's go to the canteen!"*, greeted by another friend *"Let's go!"*.[1] Such is the interaction of children when they are not able to give an idea as well as a proper answer to the question. Children will switch topics due to different vocabulary, culture, and experiences. Sometimes, children are unable to answer not because of shame or fear but due to different life experiences and the main problem is that the limited vocabulary in their mental space hinders interaction in their conversations. When describing vocabulary questions, discipline can be referred to as lexical diversity (LD). LD refers to the range and variation of unique words in a language sample and is an important indicator in assessing the breadth of vocabulary as well as the flexibility of lexical use (Jarvis, 2013; Malvern & Richards, 2002). This is important as a core construct of language development, especially in children. This is because, LD correlates with vocabulary size, semantic depth, and grammatical complexity, making it a proxy for communicative competence (cf. Burroughs, 1991; Hržica & Trtanj, 2021). Because of this, the significance of this LD shapes the communication of children, especially the marginalised groups. Increased LD occurs longitudinally reflecting cognitive maturity and integration between semantic and syntactic knowledge (see Moyle et al., 2007; Bastiaanse & Bol, 2001). The need to look at this LD is an important point to ensure that marginalised children can be measured.

[1] This situation occurred when my research colleagues and I were collecting data in one of the rural areas with status and students from hardcore poor backgrounds. This situation will apply in 2024/2025.

The measurement of LD is one thing that scholars often use to look at vocabulary ranges. As in previous studies, stating Type-token Ratio (TTR) and Number of Different Words (*NDW*) is sensitive to sample length and can result in confusing classifications in clinical contexts (see Malvern & Richards, 2002; Ratner et al., 2024). The reproducibility and testability nature of the tests makes the research to be ascertained of the correct classification level. This context is normal in research testing. The emergence of advanced measurements such as *vocabulary diversity* (VocD), MTLD (*Measure of Textual Lexical Diversity*) and MATTR (*Moving Average Type-Token Ratio*) is seen to be more stable and less influenced by text length, and shows higher psychometric validity (McCarthy & Jarvis, 2010; Kapantzoglou et al., 2019). Corpus Linguistic Statistics uses MTLD and MATTR in text measurement. Cross-language research shows that MTLD and MATTR have measurement stability across languages such as Spanish, Mandarin, and Croatian (Wu et al., 2019; Hržica & Roch, 2021). In this study, assessing the LD of marginalised children was the main focus. This is because human judgments of LD show a good correlation with computational size supporting the validity of LD constructs (Jarvis, 2017; Honko et al., 2019). The formation of the trajectory of language development is presented.

LD also increases consistently with age in a typical population and allows for comparisons of age groups as well as language ability levels (cf. Alizadeh et al., 2019; Moyle et al., 2007). Problems with vocabulary or language development often occur lately. Late talkers show a delayed but not deviant trajectory, with a less diverse vocabulary and less semantically mature (Alizadeh et al., 2019; Perry et al., 2023). This is a context that applies to describing language disorders to age variations. In Developmental Language Disorder (DLD) populations, LD scores may appear normal but lexical-semantic errors are higher indicating that LD alone is not sufficient to assess overall language function (see Charest & Skoczylas, 2019; Charest et al., 2020). The lexical-semantic relationship is an important force to create a network between words and meanings in the various contexts of the development of LD. In addition, the development of LD is also influenced by speech diversity which is a powerful resource that can be used, while printed and media inputs become more significant at the primary school level (Li et al., 2025; Montag et al., 2015). Input in language development becomes a resource for humans as language acquisition during the growth period. Contextual diversity supports the acquisition of new words and more efficient lexical processing (Norman et al., 2025; Antal et al., 2025). This also touches on the quality of inputs, not only looking at quantity but inputs play a major role in vocabulary growth (Jones & Rowland, 2017). The human input process will increase further when there is an acceptance of quality and quantity in language acquisition. Therefore, this study undertakes an important exploration of narrative data that allows measurements of LD to be tested. Narrative and expository assignments resulted in higher LD compared to typical conversation samples, demonstrating the influence of discourse context on LD scores (Spencer et al., 2023; Yu, 2010).

This study involved the narrative context of marginalised children. Various factors are also taken into account before data collection is done. Among them are social, educational, regional and economic factors. This is because the development of LD is influenced by gender factors, parenting style and socioeconomic background (Khanipour et al., 2021; Wang et al., 2022). Parenting styles and socioeconomic backgrounds are referenced through the disclosure of personal files (kept secret by school counsellors) as well as the background narration of the marginalised children themselves. Additionally, the home literacy environment is an important

mediator in vocabulary development and LD (Yu et al., 2024; Liu & Ompok, 2025). For marginalised children, they have gone through various types of literacy environments, either occurring at home or only functioning in the school library. The provision of learning spaces or literacy resources is very limited for this vulnerable group. The ecological perspective emphasizes that the development of LD is multidimensional and embedded in social and cultural contexts (Yu et al., 2024; Liu & Ompok, 2025). Through a dynamic and complex approach to modeling LD, Dynamic Systems Theory (DST) sees the development of LD as a non-linear and intraindividual variable process (Spoelman & Verspoor, 2010; van Dijk & van Geert, 2022). Variability is not considered as 'noise' but acts as a driver of development (Bashori, 2021; Sowers-Wills, 2021). Through this dynamic model allows for a more accurate understanding of longitudinal changes in LD (Zheng, 2016).

The issue of heterogeneity among marginalised children

Heterogeneity[2] among marginalised children is not a single phenomenon, but is shaped by multi-dimensional interactions such as socioeconomic status (SES), ethnicity, gender, sexual orientation, and disability (Zhao et al., 2025; Levandowski et al., 2024). An assessment of these multidimensional interactions has identified the sources that influence the lives of marginalised children. When looking at this, the concept of "intersectionality" in the context of how a person experiences discrimination or support. This approach asserts that overlapping identities produce unique patterns of risk and resilience thus reinforcing or moderating the impacts of marginalization (Coulter et al., 2021; García et al., 2023). Multidimensional factor models show that material depressants, psychosocial threats, and sociocognitive resources operate independently but interact with each other in influencing child development (DeJoseph et al., 2021). The heterogeneity perspective needs to be understood as a layered structure and not as a single category. In the context of socioeconomic inequality and the context of neighbourhood's, it is the effect of poverty that crosses the social groups of the society. Material deprivation and inequality of neighbourhood resources directly impact children's cognitive and emotional development (Ainsworth, 2002; Vargas et al., 2020). This effect directly reflects the child's willingness to accept a diverse source of environmental or environmental experiences. In addition, the gap in educational attainment is also influenced by income segregation between districts, schools as well as differences in state investment in education (Owens, 2018; Baker & Weber, 2016). Financial support is a necessity to enhance the identity of education through various levels.

In addition, cultural, linguistic and educational diversity impacts marginalised children. In the present situation, firstly, children who are not fluent in the language of the transmitter will be referred to special education because of negative interpretations that consider students sluggish or cognitively weak. Second, the way students speak loudly (culturally) is subject to disciplinary action because the wrong interpretation of the students is considered to challenge the teacher's authority. Third, the student's active learning style (kinesthetics) is referred to a specialist or counsellor because the wrong interpretation is considered to be suffering from the disorder or ADHD. These three scenarios constitute the context *of Disproportionate Referral* (Non-Proportional Referral) that occurs in cultural and linguistic diversity in the classroom

[2] Heterogeneity refers to the quality of consisting of diverse, dissimilar, or nonuniform elements, representing variability within a system, sample, or dataset.

(Novotná & Moraová, 2024; Griner & Stewart, 2013). CLD (Culturally and Linguistically Diverse) students with disabilities face multiple challenges that require culturally responsive pedagogy (Cartledge & Kourea, 2008; Williams, 2021). The role of teachers and pupils is very important to bridge the gap in teaching and learning in schools. Teacher-student relationships as well as peer support act as protective factors against the effects of discrimination and bullying in education (ten Bokkel et al., 2023; Yang et al., 2023). Through heterogeneity, education is influenced by the interaction between cultural identity and educational structures in schools.

Apart from that, the mechanism of heterogeneity involves the interaction between social pressures and sources of protection. The study of Li et al. (2021) and Wang et al. (2024), states that psychological abuse and bullying form a dynamic trajectory of social isolation and mental health problems. When a person is faced with overlapping oppression, they tend to harbor such stress, thus increasing the risk of internalizing behaviors such as depression, chronic anxiety, and self-isolation (Weiss et al., 2024; Allwood et al., 2021). This emotional burden often cannot be effectively treated due to systemic walls that make it difficult to access support services, whether due to social stigma, language barriers, or culturally insensitive bureaucracy. Family resilience and positive relationships can moderate the effects of structural stress (Zhang et al., 2022).

Further discussion of this section, looking at access to resources, opportunities and developmental outcomes of marginalised children. Heterogeneity influences access to early education, health care, and social support (Slicker et al., 2025; Mormann et al., 2025). However, there are limitations that occur, namely funding inequality and school resource gaps contribute to persistent achievement gaps (Baker & Weber, 2016; Owens, 2018). There is a possibility of uninterrupted access to the affected locality. This makes structural factors such as segregation and systematic discrimination broaden the comparison of academic outcomes and the health of marginalised groups (Henry et al., 2019; Marcus Post et al., 2025). Access and outcomes are not just individual functions but the result of systemic interactions. To look at the effectiveness of context-based policies and interventions, the effectiveness of a public policy is often compromised when it fails to address the reality of heterogeneity in populations, especially when a one-size-fits-all approach is practiced (Goldfeld et al., 2019). This rigid approach ignores differences in cultural, linguistic, and socioeconomic backgrounds, which ultimately widens the achievement gap. On the other hand, interventions centered on family involvement and cultural sensitivity, such as the ParentCorps program, have been shown to be more effective in improving students' well-being and academic achievement (Dawson-McClure et al., 2017; Hajizadeh et al., 2017). Therefore, to address this inequality holistically, the need for multi-level integrated policy responses is critical to breaking down existing systemic gaps (Unaka et al., 2024; Adeogun & Faezipour, 2023). Therefore, effective policy transformation must be culturally responsive and tailored specifically to the heterogeneity profile of the targeted community.

Use of PCA-Based Lexical Diversity Index (LDI)

In the multidimensional integration of LD, lexical diversity is a multidimensional construct that includes the abundance, variety, and distributional properties of vocabulary (Jarvis, 2013). Traditional measurements such as MTLD and Vocd-D are unidimensional and do not capture

latent structures among the various lexical features (McCarthy & Jarvis, 2010). Key Component Analysis (PCA) allows for the incorporation of various lexical variables into key components that represent latent constructs empirically (Abdi & Williams, 2010). This approach results in a more comprehensive composite index than a single measure. Based on statistics and the principle of dimension reduction, PCA converts correlated variables into orthogonal components that explain the maximum variance in the data (Jolliffe & Cadima, 2016). In this process, parameters such as eigenvalue and explained variance become critical indicators to determine which components are considered significant and stable to represent the original data structure. To ensure the consistency of the findings, the use of permutation tests such as PCA test is highly recommended to verify the statistical significance of each component selected (Camargo, 2022). This approach ensures that the index built through PCA is centred on the actual variations inherent in the data, thus avoiding errors resulting from arbitrary or random weighting.

The strength of Principal Component Analysis (PCA) lies not only in its computational capabilities, but also serves as a powerful interpretive tool through the use of biplot PCA. This technique allows visualization of the relationship between text and lexical features to be done in an easy-to-understand two-dimensional space (Kohler & Luniak, 2005). Through biplots, researchers can identify clusters, outliers, and patterns of variation that exist in a corpus more efficiently than rigid table analysis. In addition, the application of ordination techniques borrowed from the field of ecology helps to model LD as a structured system, thus providing a deeper perspective on how language units interact in a space (Ibanez, 1999; Chao & Colwell, 2022). PCA acts as a bridge that converts complex numerical data into meaningful visual representations for qualitative and quantitative analysis purposes.

The use of PCA-based indices offers significant advantages over traditional single measures, especially in terms of predictive power over higher levels of language proficiency (Woods et al., 2023; Kyle et al., 2024). In the context of second-language corpus (L2) analysis and stylometry studies, PCA has been shown to be able to distinguish the level of proficiency and writing style of individuals with greater accuracy (Craig, 2024; Park, 2024). In contrast to univariate measures that often fail to capture language nuances, this multivariate approach is much better suited to be applied in complex and heterogeneous datasets. In particular, this method is particularly relevant for studies involving wide variations in populations, such as among marginalised children, where linguistic and background factors often overlap and require a more holistic analysis.

To ensure valid and reliable PCA results, compliance with strict methodological prerequisites is critical. First, the process of standardisation and scaling of the data is a mandatory step before the analysis begins to prevent large-scale variables from unfairly dominating the variance (Syms, 2019). In addition, the stability of the resulting components is greatly influenced by the subject-to-variable ratio, where insufficient ratios can affect the effectiveness of the study results (Osborne & Costello, 2004). It should be emphasised that PCA is particularly sensitive to sample size; failure to control these parameters can result in components that are biased and do not represent the actual data structure (Elhaik, 2022; Hayashi et al., 2018). Therefore, methodological rigor is not just an option, but is also a key requirement in guaranteeing the integrity of the use of PCA as an analytical tool in complex studies.

Evolutions in statistical techniques have introduced the Robust PCA approach specifically designed to reduce the negative effects of outliers and data pollution, thus resulting in a more stable component in extreme data environments (Croux et al., 2017). In addition, the integration of PCA with machine learning algorithms as well as nonlinear techniques has expanded the capabilities of complex construct modelling, allowing nonlinear data relationships to be captured more effectively (Martin-Clemente et al., 2020; Roccatano, 2025). In this context, advanced methodological considerations are also required in dealing with the retention of small variance components to ensure that important information is not accidentally lost during the dimension reduction process (Zhang et al., 2014). The development of PCA-based LD indexes continues to evolve through computational innovations that allow language analysis to be done more dynamically and accurately.

Study Objectives

This research was conducted with the main goal of evaluating the ability of statistical integration in language measurement through three specific objectives:

- Testing whether the Moving-Average Type-Token Ratio (MATTR) and the Measure of Textual Lexical Diversity (MTLD) measures form a single, coherent construct.
- Build a robust composite index through Principal Component Analysis (PCA) to consolidate multiple lexical metrics into a single statistical representation that minimizes single measurement errors.
- Analyse the distribution of scores and rankings of students based on an index built for marking patterns of heterogeneity as well as the level of linguistic proficiency.

Data and Methodology

The study developed the *Marginalised Children's Vocabulary Index (MCVI)* as a data-driven vocabulary diversity index specifically designed to assess linguistic variation among marginalised children. This approach combines corpus linguistic analysis and multivariate statistical methods to produce stable, valid and sensitive measures of the internal heterogeneity of the population.

Measures for the Marginalised Children's Vocabulary Index (MCVI) Budget

The Marginalised Children's Vocabulary Index (MCVI) estimate is implemented through a number of systematic analytical measures to ensure that the index developed is stable, valid and based on empirical data. These steps are as follows;

Step 1: Corpus Data Preparation

Each student essay is treated as an individual document. The data is cleaned to ensure uniformity of processing, including the removal of non-linguistic symbols and the standardisation of tokenization. The length of the text is recorded as the number of tokens for the purpose of checking the stability of the index, but is not used directly in the calculation of the MCVI. The data collection protocol was carried out through the approval of the Research Ethics of the Ministry of Education (Malaysia) and the Human Ethics Evaluation Committee

of Universiti Pendidikan Sultan Idris, Malaysia (see W.A.A. Wan Halim et al., 2025). Table 1 shows the information on the background of the corpus data that has been developed. The dataset supporting the findings of this study is publicly available in the Mendeley Data repository at https://doi.org/10.17632/pybsmy8vfd.1.

Table 1. General Information of Corpus Data

Location/Zone	Age	Education	No. of Text	No. of Participants	Token
Johor	12 years old	Primary School	34	34	6,094
	16 years old	Vocational College	82	82	16,495
	18 years old	Pre-University	26	26	9,613
Perak	12 years old	Primary School	44	44	5,193
	16 years old	Vocational College	27	27	5,096
	18 years old	Pre-University	46	46	20,284
Kelantan	12 years old	Primary School	24	24	3,298
	16 years old	Vocational College	24	24	4,317
	18 years old	Pre-University	31	31	15,555
Kedah	12 years old	Primary School	35	35	3,852
	16 years old	Vocational College	30	30	6,278
	18 years old	Pre-University	22	22	13,261
Sabah	12 years old	Primary School	22	22	2,673
	16 years old	Vocational College	64	64	13,188
	18 years old	Pre-University	70	70	28,421
Sarawak	12 years old	Primary School	30	30	3,084
	16 years old	Vocational College	30	30	5,546
	18 years old	Pre-University	29	29	14,779
		Total	670	670	177,027

Step 2: Selection of Lexical Diversity Measures

Two more stable measures of LD were chosen, namely the Moving Average Type–Token Ratio (MATTR50) and the Measure of Textual Lexical Diversity (MTLD). The Type–Token Ratio (TTR) measure is not used due to the high sensitivity to the length of the text, which can result in misleading interpretations. This study used #LancsBoxX software founded by Brezina (2025) for MATTR50 and MTLD assessments.

Step 3: Standardise scores

The MATTR50 and MTLD scores are converted to standard scores (z-scores) to ensure that both variables are on the same scale. This standardisation avoids the dominance of any measure in subsequent analyses and is an important prerequisite for multivariate analysis.

Step 4: Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was conducted on MATTR50 and MTLD standard scores to identify the latent structure of LD. The first component (PC1), which describes the highest percentage of variance, is interpreted as a latent construct of vocabulary diversity. The component load is used as an empirical weighting in the construction of the index.

Step 5: Establishment of the Lexical Diversity Index (LDI)

The PC1 score is extracted and interpreted as the Lexical Diversity Index (LDI), which is a composite measure that combines local lexical variation and the continuity of global variation in a balanced and data-driven manner.

Step 6: Conversion to Index Scale (MCVI)

For ease of interpretation and comparison, the LDI score is normalised to a scale of 0–100. This score is then referred to as the Marginalised Children's Vocabulary Index (MCVI), which represents the relative position of each student in the study cohort.

Step 7: Classification and Interpretation

MCVI scores were analysed using standard deviation to classify the level of vocabulary diversity into the categories of high, medium high, medium low and low. This classification allows the analysis of internal linguistic heterogeneity and the profiling of students to be carried out systematically.

Calculation of Marginalised Children's Vocabulary Index (MCVI)

Primary School Data (SR)

The analysis focused on two more stable measures, namely MATTR50 (*local variation*) and MTLD (*global sustainability*).

Z-score standards

Metrics are standardised because MATTR50 and MTLD are on different scales;

$$z_X = \frac{X - \mu_X}{\sigma_X}$$

Indicators:
X = student score (e.g. student MATTR50 A)
μ_x = Mean (average) score for all students in the cohort
σ_x = Standard Deviation (SD) score for all cohort students

For example;
MATTR50 students A = **0.80**
Average cohort MATTR50, μ\muμ = **0.72**
Standard deviation MATTR50 cohort, σ\sigmaσ = **0.06**

Cast:

$$z_{MATTR50} = \frac{0.80 - 0.72}{0.06} = \frac{0.08}{0.06} = 1.33$$

So, students A 1.33 SD is above average for MATTR50.

PCA to build latent constructs of Lexical Diversity

PCA builds the first component (PC1) as a combination of liners:

$$PC1 = a(z_{MATTR50}) + b\ (z_{MTLD})$$

Verdict PCA (SR)
Explained variance (PC1) = 0.9257 (≈92.57%)
Loadings;
α = 0.7071 for MATTR50
b = 0.7071 for MTLD

Cast:

$$LDI_{PC1} = 0.7071\ (z_{MATTR50}) + 0.7071\ (z_{MTLD})$$

Both measures contribute in a balanced manner, so the index is data-driven and not arbitrary weights.

Conversion of PC1 score to a scale of 0-100 (MCVI_100)

Practical interpretation and ranking:

$$MCVI_{100} = 100 \times \frac{LDI_{PC1} - LDI_{min}}{LDI_{max} - LDI_{min}}$$

In SR
MCVI_100 min = 0
MCVI-100 max = 100
min ≈ 43.74, SD≈ 17.33

First Output: MCVI Ranking (SR)

Tables 2 and 3 are the results for MCVI which is ranked 10 places above and 10 places below. In this MCVI, a scale of 1-100 is used to measure the ranking of each student.

Table 2. Top 10 (MCVI_100 highest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
1	SRB018Lii.txt	102	0.86	112.04	100.00
2	SRG005.txt	177	0.82	105.80	91.69
3	SRS026.txt	92	0.81	98.75	86.85
4	SKP-28.txt	201	0.79	97.59	83.68
5	SRB003Piv.txt	268	0.80	93.77	83.05
6	SRG029.txt	85	0.81	85.00	79.92
7	SKP-3.txt	189	0.79	89.89	79.80
8	SRS002.txt	81	0.79	87.48	78.59
9	SKP-5.txt	202	0.80	82.38	77.31
10	SRT011.txt	109	0.78	86.29	76.70

Table 3. Bottom 10 (MCVI_100 lowest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
187	SRT013.txt	210	0.46	16.07	0.00
186	SRG020.txt	82	0.49	21.25	6.48

185	SRS027.txt	93	0.52	14.67	7.04
184	SRS023.txt	108	0.53	19.08	10.56
183	SRKM024.txt	97	0.55	14.07	10.62
182	SRKM017.txt	106	0.56	19.64	14.71
181	SRS018.txt	122	0.55	22.37	14.80
180	SRT002.txt	86	0.55	23.20	15.21
179	SRS016.txt	113	0.55	26.11	16.68
178	SRB009Piv.txt	108	0.57	21.48	16.93

Table 4 shows the distribution for the categories that use the measurement of MCVI scores, the number of texts and the percentage of students who are in 4 different positions, namely high, medium high, medium low and low.

Table 4. The distribution of categories by SD for MCVI_z is the result of SR (N=189)

Interpretation	Score Range MCVI (z-score)	No. Text	Percent
Height	> +1 SD	27	14.29
Medium height	0 up to +1 SD	64	33.86
Medium low	-1 up to 0 SD	64	33.86
Low	< -1 SD	34	17.99

Expectations:
Students > +1 SD are well above average
Students < −1 SD are well below average

Data Vocational College (KV)

The MCVI calculation is also used to form the index and ranking of each marginalised child for vocational college (KV) data.

Second Output: Rangking MCVI (KV)

Tables 5 and 6 show the position of vocational college (KV) data for marginalised children according to the index which belongs to 2 divisions, namely the top 10 and the bottom 10.

Table 5. Top 10 (MCVI_100 highest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
1	SVB027Piii.txt	270	0.88	166.74	100.00
2	SVT021.txt	185	0.85	134.45	82.88
3	KVPGS2P-51.txt	161	0.84	121.65	76.43
4	SVB021Piii.txt	84	0.85	116.22	76.18
5	KVPGS2P-13.txt	234	0.83	120.41	74.22
6	KVPGS2P-49.txt	186	0.84	112.27	72.98
7	KVPGS1P-27.txt	276	0.82	115.57	70.69
8	KVPGS2P-27.txt	185	0.83	110.55	70.60
9	SVB038Liii.txt	239	0.84	105.43	70.47
10	KVG014.txt	209	0.82	114.28	70.22

Table 6. Bottom 10 (MCVI_100 lowest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
257	SVB044Liii.txt	188	0.59	32.80	0.00
256	SVB046Liv.txt	165	0.63	33.34	7.20
255	SVT014.txt	120	0.63	35.00	7.81
254	SVB004Liv.txt	109	0.67	26.53	11.71
253	SVB041Liv.txt	225	0.67	31.64	13.58
252	SVS017.txt	216	0.67	32.74	13.99
251	SVKM020.txt	194	0.66	41.74	15.54
250	SVT011.txt	118	0.67	40.44	16.82
249	SVB002Liv.txt	165	0.68	39.16	18.10
248	SVKM026.txt	190	0.69	38.00	19.42

Table 7 shows the MCVI category distribution of Vocational College (KV) data based on SD by measuring 257 marginalised children texts.

Table 7. MCVI category distribution of Vocational College (KV) data based on SD through N=257

Interpretation	Score Range MCVI (z-score)	No. Text	Percent
Height	> +1 SD	34	13.23
Medium height	0 up to +1 SD	87	33.85
Medium low	-1 up to 0 SD	96	37.35
Low	< -1 SD	40	15.56

Expectations:
Students > +1 SD are well above average
Students < −1 SD are well below average

Data Pre-University

Next, MCVI analysis was conducted on pre-university data involving 224 texts by marginalised children.

Third Output: Ranking MCVI (PRAU)

Tables 8 and 9 show the results of the MCVI ranking of pre-university data, the table presents information with the ladder of 10 and 10 and 10 below.

Table 8. Top 10 (MCVI_100 highest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
1	SABP24.txt	330	0.87	181.01	100.00
2	SMG035.txt	500	0.86	153.90	87.93
3	SMG007.txt	281	0.85	148.17	83.62
4	SMB048Liii.txt	250	0.86	132.66	80.22
5	SABP26.txt	268	0.85	128.51	76.48
6	SABP22.txt	532	0.85	127.36	76.07

7	SMS016.txt	424	0.84	125.37	73.12
8	SABP12.txt	379	0.83	125.18	70.82
9	SABP2.txt	235	0.85	111.76	70.40
10	SABP3.txt	419	0.84	114.62	69.21

Table 9. Bottom 10 (MCVI_100 lowest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
224	SMB036Lii.txt	581	0.66	34.32	0.00
223	SMG003.txt	363	0.69	36.62	7.51
222	SMB044Piii.txt	299	0.68	43.21	7.68
221	SABP11.txt	250	0.70	47.41	13.65
220	SMB040Piii.txt	240	0.70	52.45	15.48
219	SMB003Piii.txt	289	0.72	47.88	18.27
218	SABP7.txt	223	0.72	48.01	18.32
217	SMS017.txt	402	0.73	43.61	18.95
216	SMS024.txt	466	0.72	51.46	19.57
215	SMKM021.txt	513	0.72	51.76	19.68

Table 10, showing the distribution of pre-university data, is a high percentage of 37.05% with a total of 83 texts that are at medium low.

Table 10. Distribution of MCVI category (PRAU) based SD through (N=224)

Interpretation	Score Range MCVI (z-score)	No. Text	Percent
Height	> +1 SD	37	16.52
Medium height	0 up to +1 SD	70	31.25
Medium low	-1 up to 0 SD	83	37.05
Low	< -1 SD	34	15.18

Expectations:
Students > +1 SD are well above average
Students < −1 SD are well below average

Data Overall MCVI

The following is the overall data for MCVI by ranking involving the three levels of education, namely Primary School (SR), Vocational College (KV) and Pre-University. Tables 11 and 12 show the 10 above and 10 below for the MCVI_100 ranking.

Table 11. Top 10 (MCVI_100 highest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
1	SABP24.txt	330	0.87	181.01	100.00
2	SVB027Piii.txt	270	0.88	166.74	96.71
3	SMG035.txt	500	0.86	153.90	90.33
4	SMG007.txt	281	0.85	148.17	87.35
5	SMB048Liii.txt	250	0.86	132.66	83.67

6	SVT021.txt	185	0.85	134.45	83.05
7	SABP26.txt	268	0.85	128.51	81.19
8	SABP22.txt	532	0.85	127.36	80.83
9	SMS016.txt	424	0.84	125.37	79.03
10	KVPGS2P-51.txt	161	0.84	121.65	77.86

Table 12. Bottom 10 (MCVI_100 lowest)

Rank	Participants	Tokens	MATTR50	MTLD	MCVI_100
668	SRT013.txt	210	0.46	16.07	0.00
667	SRG020.txt	82	0.49	21.25	5.16
666	SRS027.txt	93	0.52	14.67	6.63
665	SRS023.txt	108	0.53	19.08	9.19
664	SRKM024.txt	97	0.55	14.07	9.98
663	SRS018.txt	122	0.55	22.37	12.58
662	SRT002.txt	86	0.55	23.20	12.84
661	SRKM017.txt	106	0.56	19.64	12.90
660	SRS016.txt	113	0.55	26.11	13.75
659	SRB009Piv.txt	108	0.57	21.48	14.66

Table 13 shows the distribution of the overall data by showing that 36.72% of MCVI is at the Medium High level and 32.99% is at the Medium Low level. The high and low values are at a balanced level.

Table 13. MCVI Category Distribution (Overall, SD-based)

Interpretation	Score Range MCVI (z-score)	No. Text	Percent
Height	> +1 SD	102	15.22
Medium height	0 up to +1 SD	246	36.72
Medium low	-1 up to 0 SD	221	32.99
Low	< -1 SD	101	15.07

Expectations:
Students > +1 elementary school are well above average
Students < −1 elementary school are well below average

Findings

This section presents the findings of the analysis of the Marginalised Children's Vocabulary Index (MCVI) based on the entire corpus as well as the breakdown by level of education, namely primary school (SR), vocational college (KV) and pre-university (PRAU). The main focus of the findings was on the distribution of the index, its stability to the length of the text, as well as the degree of internal linguistic heterogeneity in the population of marginalised children.

Overall Distribution of MCVI

Analysis of a total of 670 essays showed that MCVI scores were scattered almost normally with concentrations in the mid-range of the scale. The majority of students are in the medium category, while the percentage of students in the high and low categories is almost balanced. This pattern suggests that vocabulary diversity among marginalised children is not uniform, but rather shows significant internal variation despite being in relatively similar socioeconomic backgrounds. This stable distribution also confirms that MCVI serves as a meaningful index for linguistic profiling rather than just a statistical artefact.

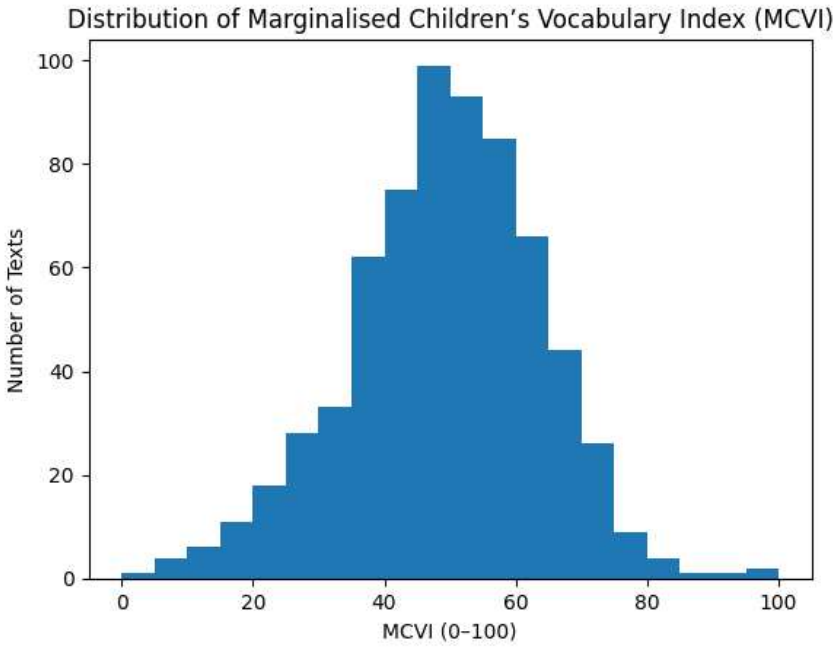


Figure 1. Distribution of the Marginalised Children's Vocabulary Index (MCVI)

Figure 1 of the Marginalised Children's Vocabulary Index (MCVI) distribution shows the distribution of vocabulary diversity scores for a total of 670 essays by marginalised children. MCVI scores span the entire index range from 0 to 100, with an average (mean) value of 49.25 and a standard deviation of 14.53, indicating a moderate but significant degree of variation in students' lexical ability. In terms of the shape of the sprinkle, the skewness value −0.21 indicates a slight tilt to the left, but the magnitude is low and indicates an almost symmetrical sprinkle. Meanwhile, a kurtosis value of 0.35 (excess kurtosis) suggests a slightly more pointed distribution than a normal distribution, but still within a range that can be considered moderate and not extreme. This combination of skewness and kurtosis values shows that the distribution of MCVI is not predominantly influenced by the presence of outliers. In terms of classification based on standard deviation, the majority of students fall into the moderately high (36.72%) and moderately low (32.99%) categories, while the high (15.22%) and low (15.07%) categories form a nearly balanced minority group. Overall, these statistical features confirm that MCVI has a stable distribution nature and is suitable for use as a quantitative measure to reflect internal linguistic heterogeneity among marginalised children.

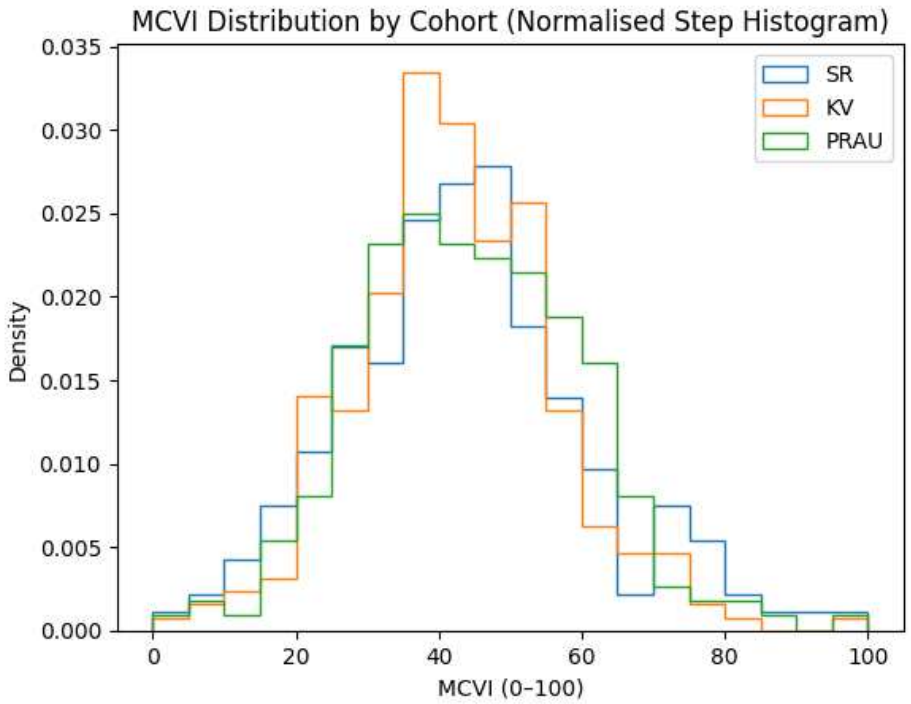


Figure 2. Distribution of Marginalised Children's Vocabulary Index (MCVI) by Education Cohort (Normalised Step Histogram)

Figure 2, the distribution of MCVI by educational cohort presented in the form of a normalised step histogram allows the comparison of the distribution of vocabulary diversity between primary schools (SR), vocational colleges (KV) and pre-universities (PRAU) to be done relatively without being influenced by differences in sample size. In terms of descriptive statistics, the average MCVI for all three cohorts was in the mid-range of the scale, which was around SR ≈ 47, KV ≈ 43, and PRAU ≈ 44, with a relatively large standard deviation (around 14–15 index points) for each cohort. This almost identical standard deviation value indicates that the level of internal variation is consistent across levels of education. From the point of view of standard-deviation-based classification, the primary school cohort showed a slightly higher percentage of students in the lower category (≈18%) than in the high category (≈14%), reflecting the still uneven development of vocabulary. Within the vocational college cohort, the majority of students were concentrated in the moderately low (≈37%) and moderately high (≈34%) categories, with the smaller extreme categories, signaling a more concentrated distribution around mid-scores. On the other hand, the pre-university cohort showed a relative increase in the high category (≈17%), but the low category still remained around 15%, suggesting that the increase in education level did not eliminate internal linguistic heterogeneity.

The Relationship Between MCVI and Text Length

To assess the stability of MCVI against the length of the text, the relationship between the MCVI score and the number of tokens was analysed. The findings show that the relationship between the two variables is weak, with a wide point distribution and the absence of a clear linear pattern. These results confirm that MCVI does not automatically increase with the increase in text length, thus supporting the rationale for MATTR50 size selection and MTLD

which are relatively independent of the effect of text length. Compared to the Type–Token Ratio (TTR), MCVI was found to be more stable and suitable for use in the context of student essays of varying lengths.

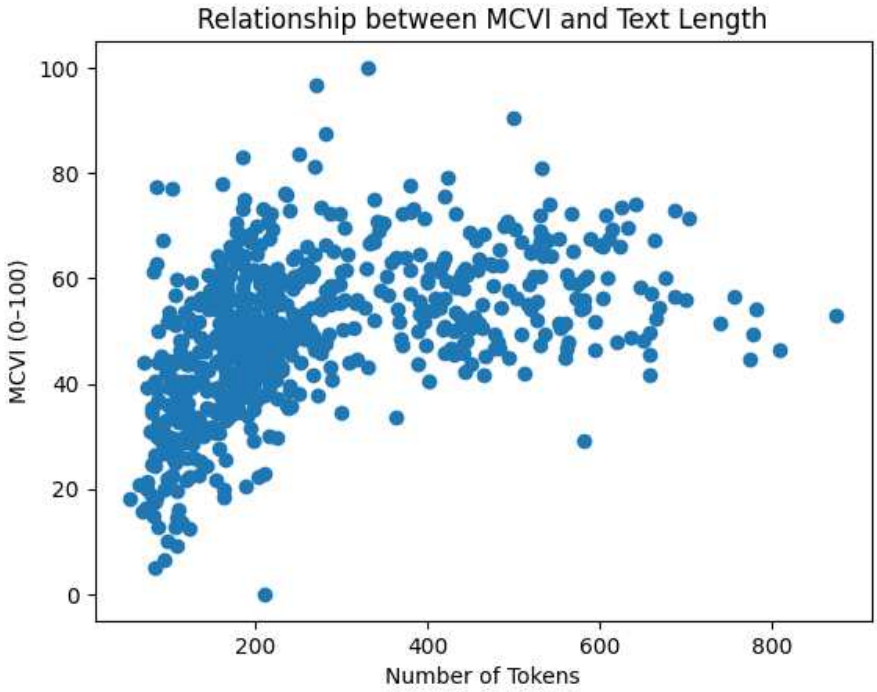


Figure 3. Relationship between MCVI and Text Length (Number of Tokens)

Figure 3, the relationship between the Marginalised Children's Vocabulary Index (MCVI) and text length measured through the number of tokens shows that vocabulary diversity does not depend directly on the quantity of text produced. Overall, the point distribution in the graph shows a wide spread without a clear linear pattern, with a simple Pearson correlation coefficient value, which is $r = 0.46$. While there is a tendency for MCVI to increase as the number of tokens increases, the strength of this relationship is limited and does not reflect strong dependence. This suggests that students with longer texts do not necessarily result in higher vocabulary diversity, and instead students with shorter texts still have the potential to achieve high MCVI scores.

When these relationships were analysed by education level, the strength of the correlation was found to be decreasing. For primary schools, the relationship between MCVI and text length was weak to moderate ($r \approx 0.33$), reflecting that in the early stages of language development, vocabulary diversity was still slightly influenced by the amount of linguistic production. However, for vocational colleges, the correlation became weaker ($r \approx 0.16$), while at the pre-university level, the correlation was close to zero ($r \approx 0.00$). This pattern suggests that as education levels increase, the diversity of students' vocabulary increasingly reflects actual linguistic ability rather than mere text quantity. Overall, these findings confirm that MCVI is a relatively independent index of text length and more stable than traditional measures such as the Type–Token Ratio (TTR), thus supporting its suitability for student essay analysis with significant variations in text length.

Comparison of MCVI Distribution by Education Level

The next analysis compared the distribution of MCVI by education level. For primary schools, the distribution of MCVI shows a wide range of variation with student attendance at both ends of the spectrum, reflecting the still uneven level of lexical development. Among vocational college students, the distribution of MCVI is more concentrated around the middle score, yet internal variation is still significant. Meanwhile, for pre-university, despite an increase in the number of students in the high MCVI category, the range of variation remains wide, signifying that academic maturity does not automatically result in homogeneous vocabulary diversity.

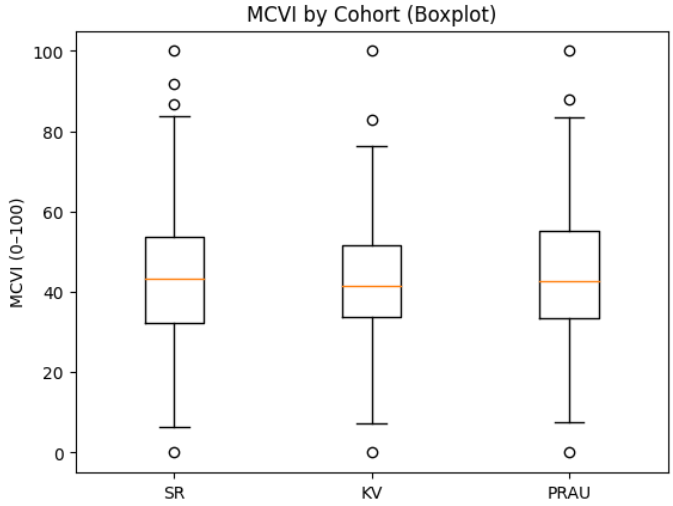


Figure 4. Distribution of MCVI by Educational Level (Boxplot)

Figure 4, the MCVI boxplot graph by education cohort shows a comparison of the distribution of vocabulary diversity for three levels of education, namely primary school (SR), vocational college (KV) and pre-university (PRAU). In general, all three cohorts showed a median MCVI that was in the mid-range of the index scale, indicating that the average difference in vocabulary diversity between education levels was moderate. However, the boxplot analysis revealed that the interquartile range (IQR) of each cohort was relatively wide, particularly among primary and pre-university students, reflecting a significant level of internal linguistic heterogeneity.

For the primary school cohort, the MCVI distribution showed the widest range with the presence of values at both the low and high ends of the spectrum. This pattern reflects the variation in language development that is still uneven in the early stages of schooling. Among vocational college students, the distribution of MCVI was found to be slightly more concentrated around the median, indicating a more uniform level of vocabulary diversity compared to primary school, yet still maintaining significant internal variation. Meanwhile, for the pre-university cohort, although the median MCVI was slightly higher, the large interquartile range suggests that increased levels of education do not necessarily result in homogeneous vocabulary diversity among students.

The absence of extreme median differences between the three cohorts, combined with wide internal variations, suggests that individual factors and personal linguistic experiences play a more dominant role than education level factors alone. Thus, this graph reinforces the argument that MCVI is more effective in revealing internal linguistic heterogeneity than in roughly distinguishing groups based on education level.

Comparison of Mean MCVI and Internal Heterogeneity

A mean MCVI analysis with an error of ±1 standard deviation provides a synthesis picture of the differences between cohorts. The mean MCVI for all three levels of education was in almost the same range, indicating that the average difference in vocabulary diversity between primary schools, vocational colleges and pre-universities was limited. However, large standard deviations per cohort suggest that internal variations are much more significant than differences between levels of education.

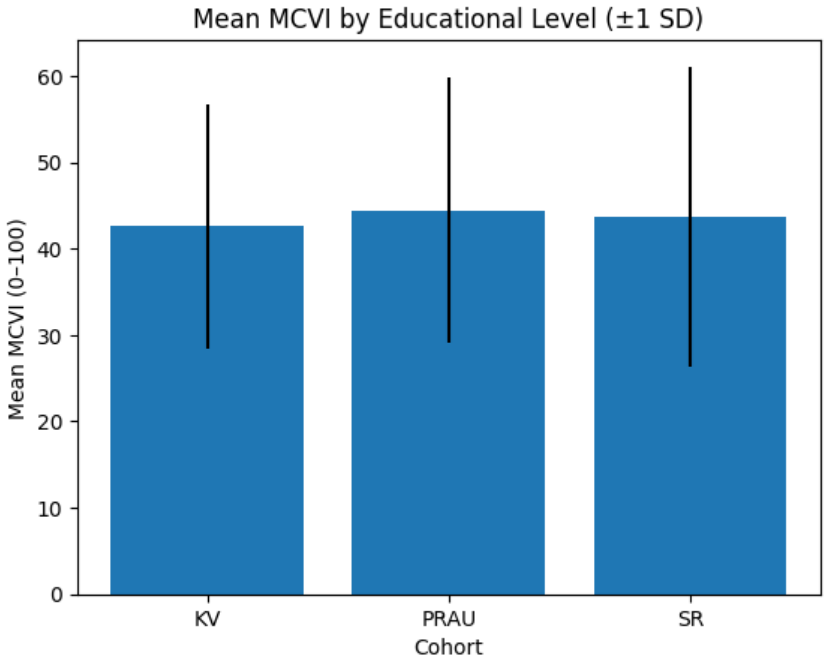


Figure 5. Mean MCVI by Educational Level with ±1 Standard Deviation

Figure 5, MCVI Mean Graph by Educational Level (±1 SD) shows a comparison of the average value of students' vocabulary diversity across three levels of education, namely primary school (SR), vocational college (KV) and pre-university (PRAU), along with the error ±1 standard deviation as an indicator of internal variation. Statistically, the average MCVI score for primary schools is around 47.0 with a standard deviation of about 14.8, while vocational colleges record a slightly lower average, which is around 42.6 with a standard deviation of 14.1. For pre-university, the average MCVI is around 44.4 with a standard deviation of 15.3. This average range shows that the mean difference in MCVI between education levels is relatively small, i.e. less than 5 index points, and remains in the mid-range of the MCVI scale.

On the other hand, the large and almost equivalent standard deviation values of all three cohorts (around 14–15 index points) indicate that the internal variation within each cohort is high. For example, for elementary school cohorts, the standard deviation range ±1 includes MCVI scores ranging from 32 to 62, while for pre-university the range is around 29 to 60. This means that within each level of education, although the average difference in MCVI between primary ($M \approx 47.0$), vocational college ($M \approx 42.6$) and pre-university ($M \approx 44.4$) is small, the large and nearly equivalent standard deviation ($SD \approx 14–15$) suggests that the internal variation in each cohort is much more significant than the difference between levels of education.

Discussion and Summary

Overall MCVI Distribution and Cohort Distribution

The almost normal overall distribution of MCVI with the dominance of moderate categories supports the view that LD is a core construct but not homogeneous in language development. These findings are in line with previous studies that showed that LD correlates with vocabulary size, semantic depth, and grammatical complexity, yet develops unevenly between individuals (Burroughs, 1991; Jarvis, 2013; Hržica & Trtanj, 2021). The almost balanced percentage between the high and low categories reflects that although children are in the same marginalised context, their language development trajectories are layered and varied. When analysed by educational cohort, normalised step histogram findings showed that primary schools had the widest distribution, in line with longitudinal findings that the increase in LD in the early stages was strongly influenced by cognitive maturity and unstable semantic-syntactic integration (Moyle et al., 2007; Bastiaanse & Bol, 2001). The shift in distribution towards higher scores in pre-university supports the finding that LD increases with age, but does not necessarily become homogeneous (Alizadeh et al., 2019). Conceptually, these findings reinforce the argument that heterogeneity is a structural feature of language development, not an unusual deviation.

The Relationship Between MCVI and Text Length

The weak relationship between MCVI and text length provides strong empirical support for the methodological critique of the Type–Token Ratio (TTR). As Malvern and Richards (2002) and Jarvis (2013) argue, TTR fails to distinguish between actual lexical variations and text-length mathematical artifacts. The findings of this study show that when MATTR50 and MTLD were combined through the PCA approach, the influence of text length was reduced and almost zero at the pre-university level. This is in line with evidence that advanced measures such as MTLD and MATTR are more stable and have higher psychometric validity across text and language lengths (McCarthy & Jarvis, 2010; Wu et al., 2019; Hržica & Roch, 2021). Theoretically, the decline in the MCVI–Tokens correlation by educational level supports the view that as language systems become more mature, vocabulary diversity reflects actual linguistic competence and not the quantity of language production (Jarvis, 2017; Honko et al., 2019). This confirms the MCVI as a more defensible index compared to traditional unidimensional measurements.

Comparison of MCVI Distribution by Education Level

The MCVI boxplot shows that the median of vocabulary diversity is relatively close between cohorts, while the internal variation within each cohort is wide. These findings challenge the linear assumption that increased levels of education will result in homogeneous LD. Rather, it aligns with studies showing that although LD increases on average with age, individual trajectories remain different, including in typical and a-typical populations such as late talkers and DLD (Alizadeh et al., 2019; Perry et al., 2023; Charest & Skoczylas, 2019). From an ecological and intersectionality perspective, these internal variations can be understood as the result of complex interactions between linguistic inputs, social contexts, literacy environments,

and socioeconomic factors (Jones & Rowland, 2017; Yu et al., 2024; Zhao et al., 2025). Thus, these findings reinforce the argument that linguistic heterogeneity is not just a random individual variation, but a manifestation of layered structures in marginalised children's language development.

Comparison of Mean MCVI and Dominance of Internal Heterogeneity

The small mean differences in MCVI between primary schools, vocational colleges and pre-universities, combined with large and almost equivalent standard deviations, suggest that internal variations are much more dominant than differences between education levels. These findings are in line with the Dynamic Systems Theory approach which sees variability not as "noise", but as a key driver of developmental change (Spoelman & Verspoor, 2010; van Dijk & van Geert, 2022). In this context, MCVI variability reflects an ever-changing language system, sensitive to individual input and experience. From the point of view of educational and clinical implications, these findings support the suggestion that advanced measures such as MTLD and MATTR are more suitable for progress monitoring and intervention planning than single measures (Kapantzoglou et al., 2019; McCarthy & Jarvis, 2010). However, in parallel with Charest et al. (2020), LD needs to be combined with error analysis and semantic depth for a thorough language assessment.

Rujukan

Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433–459.

Adeogun, A., & Faezipour, M. (2023). Data-driven approaches to support marginalised children: Toward integrated multi-level interventions. *Journal of Social Policy and Society*.

Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces, 81*(1), 117–152.

Alizadeh, M., Soleymani, Z., Jalaie, S., Shahrivar, N. M., et al. (2019). Lexical diversity in Persian late talkers: A pilot study. *Koomesh*.

Allwood, M. A., Bell, D. J., & Husain, S. A. (2021). Psychological distress and discrimination among marginalized youth: A developmental perspective. *Journal of Youth and Adolescence, 50*(4), 689–704.

Antal, C., Johns, B. T., & de Almeida, R. G. (2025). The effect of contextual and semantic diversity in lexical and conceptual access: Evidence from a picture–word semantic congruency task. *Language, Cognition and Neuroscience*.

Assari, S., & Caldwell, C. H. (2019). Family income at birth and risk of attention deficit hyperactivity disorder at age 15: Race by gender differences. *Children, 6*(1), 10.

Baker, B. D., & Weber, M. A. (2016). State school finance inequities and educational disparities. *Educational Policy, 30*(2), 237–265.

Bashori, M. (2021). The development of intra-individual variability in academic writing: A study on lexical diversity and lexical sophistication. *Studies in English Language and Education*.

Bastardas-Boada, A. (2013). Sociolinguistics: Towards a complex ecological view. In *Understanding Complex Systems*. Springer.

Bastiaanse, R., & Bol, G. (2001). Verb inflection and verb diversity in three populations: Agrammatic speakers, normally developing children, and children with specific language impairment. *Brain and Language*.

Burroughs, E. (1991). Lexical diversity in listeners' judgments of children. *Perceptual and Motor Skills*.

Brezina, V. (2025). Corpus linguistics and AI: #LancsBox X in the context of emerging technologies. *International Journal of Language Studies*, *19*(2), 75–90. https://doi.org/10.5281/zenodo.15250820

Camargo, A. (2022). PCAtest: Testing the statistical significance of principal component analysis in R. *PeerJ, 10*, e12967.

Cartledge, G., & Kourea, L. (2008). Culturally responsive classrooms for culturally diverse students with and at risk for disabilities. *Exceptional Children, 74*(3), 351–371.

Chao, A., & Colwell, R. K. (2022). Biodiversity: Concepts, dimensions, and measures. In *The Ecological and Societal Consequences of Biodiversity Loss*.

Charest, M., & Skoczylas, M. J. (2019). Lexical diversity versus lexical error in the language transcripts of children with developmental language disorder: Different conclusions about lexical ability. *American Journal of Speech-Language Pathology*.

Charest, M., Skoczylas, M. J., & Schneider, P. (2020). Properties of lexical diversity in the narratives of children with typical language development and developmental language disorder. *American Journal of Speech-Language Pathology*.

Coulter, R. W. S., Kenst, K. S., Bowen, D. J., & Scout. (2021). Research funded by the National Institutes of Health on the health of lesbian, gay, bisexual, and transgender populations. *American Journal of Public Health, 104*(2), e105–e112.

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.

Craig, H. (2024). Principal components analysis in stylometry. *Digital Scholarship in the Humanities*.

Croux, C., Garcia-Escudero, L. A., Gordaliza, A., & Martin, R. S. (2017). Robust principal component analysis based on trimming around affine subspaces. *Statistica Sinica, 27*(3), 1437–1459.

Cruz, R. A., Firestone, A. R., & Rodl, J. E. (2025). Educational access and inequities among rural marginalized youth. *Rural Education Quarterly*.

Dawson-McClure, S., Calzada, E. J., Huang, K.-Y., Kamboukos, D., Rhule, D., Kolawole, B., Petkova, E., & Brotman, L. M. (2017). A population-level approach to promoting healthy child development and school success in low-income, urban neighborhoods. *Journal of Child and Family Studies, 26*(7), 1964–1976.

DeJoseph, M. L., Sifre, R., Raver, C. C., & Blair, C. (2021). Multidimensional poverty and child development outcomes: A latent variable approach. *Developmental Psychology, 57*(4), 532–546.

Elek, C., Pink, B., & Dahlstrom, J. (2020). Educational engagement among Indigenous children in remote communities. *Australian Journal of Education, 64*(3), 245–262.

Elhaik, E. (2022). Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Scientific Reports, 12*, 14683.

García, J. J., Sharif, M. Z., & Ompad, D. C. (2023). Intersectionality and child health inequities. *Social Science & Medicine, 317*, 115612.

Goldfeld, S., O'Connor, E., & Sayers, M. (2019). Precision public health: Using data to better target interventions for children. *The Lancet Public Health, 4*(6), e275–e276.

Griner, A. C., & Stewart, M. L. (2013). Addressing disproportionality in special education. *Educational Researcher, 42*(6), 375–384.

Hajizadeh, N., Stevens, E., & Pinderhughes, E. (2017). Culturally grounded interventions for marginalized youth. *Journal of Community Psychology, 45*(5), 617–633.

Henry, D. B., Tolan, P. H., & Gorman-Smith, D. (2019). Structural inequality and child outcomes: A longitudinal community-based study. *American Journal of Community Psychology, 63*(1–2), 55–69.

Honko, M., Jarvis, S., & Vainio, S. (2019). Readers' perceptions of lexical diversity: Examining lexical diversity at the interface between quantitative and qualitative research. *Virittäjä*.

Hržica, G., & Roch, M. (2021). Lexical diversity in bilingual speakers of Croatian and Italian. *Trends in Language Acquisition Research*.

Hržica, G., & Trtanj, I. (2021). Measures of lexical diversity in narratives of preschool and early school age children. *Rasprave Instituta za Hrvatski Jezik i Jezikoslovlje*.

Ibanez, F. (1999). Diversity and ordination. *Oceanis, 25*(3–4), 121–143.

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

Jarvis, S. (2017). Grounding lexical diversity in human judgments. *Language Testing*.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A, 374*, 20150202.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and diversity of the input on vocabulary growth. *Cognitive Psychology*.

Kapantzoglou, M., Fergadiotis, G., & Buenavides, A. A. (2019). Psychometric evaluation of lexical diversity indices in Spanish narrative samples from children with and without developmental language disorder. *Journal of Speech, Language, and Hearing Research*.

Khanipour, H., Aghazamani, M., Nezafatbaksh, S., & Tarvand, S. (2021). The influence of child gender, parents' job and education, number of siblings and parenting styles on child development. *Language Related Research, 12*(3), 45–67.

Kohler, U., & Luniak, M. (2005). Data inspection using biplots. *Stata Journal, 5*(2), 208–223.

Kramsch, C. (2008). Ecological perspectives on foreign language education. *Language Teaching*.

Kyle, K., Sung, H., Eguchi, M., & Zenker, F. (2024). Evaluating evidence for the reliability and validity of lexical diversity indices in L2 oral task responses. *Studies in Second Language Acquisition*.

Levandowski, B. A., Stansfield, K., & Saito, K. (2024). Intersectional identities and mental health disparities among marginalized children. *Journal of Child Psychology and Psychiatry, 65*(2), 214–227.

Li, L., Song, M., & Cai, Q. (2025). The words children hear and see: Lexical diversity across modalities and its impact on lexical development. *Developmental Science*.

Li, X., Wang, L., & Chen, J. (2021). Bullying victimization and psychological distress among marginalized adolescents. *Journal of Adolescence, 88*, 87–98.

Liu, Q., & Ompok, C. S. C. (2025). The influence of parenting styles and child–parent relationship on home literacy environment: Empirical evidence from Malaysia. *Humanities and Social Sciences Letters*.

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

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

Marcus Post, M., Schaeffer, C., & Rivera, L. (2025). Systemic barriers and educational inequity. *Educational Evaluation and Policy Analysis*.

Martin-Clemente, R., Zarzoso, V., & Camargo-Olivares, J. L. (2020). On the discriminative properties of principal component analysis based on L1-norm. *Proceedings of the International Conference on Computational Science and Computational Intelligence*.

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.

Montag, J. L., Jones, M. N., & Smith, L. B. (2015). The words children hear: Picture books and the statistics for language learning. *Psychological Science*.

Mormann, E., Steinmayr, R., & Spinath, B. (2025). Socioeconomic inequality and academic achievement trajectories. *Learning and Instruction*.

Moyle, M. J., Weismer, S. E., Evans, J. L., & Lindstrom, M. J. (2007). Longitudinal relationships between lexical and grammatical development in typical and late-talking children. *Journal of Speech, Language, and Hearing Research*.

Norman, R., Taylor, J. S. H., & Rodd, J. M. (2025). The effects of contextual diversity on lexical processing: A scoping review. *Psychonomic Bulletin & Review*.

Novotná, K., & Moraová, H. (2024). Cultural and linguistic diversity and special education referral. *European Journal of Special Needs Education, 39*(1), 15–29.

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject-to-item ratio in principal components analysis. *Practical Assessment, Research and Evaluation, 9*(11).

Owens, A. (2018). Income segregation between school districts and inequality in academic achievement. *Sociology of Education, 91*(1), 1–27.

Park, S. (2024). Identifying key linguistic variables of second language speaking proficiency using principal component analysis. *Forum for Linguistic Studies*.

Perry, L. K., Kucker, S. C., Horst, J. S., & Samuelson, L. K. (2023). Late bloomer or language disorder? Differences in toddler vocabulary composition associated with long-term language outcomes. *Developmental Science*.

Ratner, N. B., Han, Y., & Yang, J. S. (2024). Should we stop using lexical diversity measures in children's language sample analysis? *American Journal of Speech-Language Pathology*.

Roccatano, D. (2025). Principal component analysis of molecular dynamic trajectories: Concepts, tools, and applications. *Wiley Interdisciplinary Reviews: Computational Molecular Science*.

Slicker, G., Kim, J., & Choi, S. (2025). Early childhood intervention access and socioeconomic disparities. *Early Childhood Research Quarterly*.

Spencer, T. D., Tolentino, T. J., & Foster, M. E. (2023). Impact of discourse type and elicitation task on language sampling outcomes. *American Journal of Speech-Language Pathology*.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*.

Syms, C. (2019). Principal components analysis. In *Encyclopedia of Ecology* (2nd ed.).

ten Bokkel, I. M., Verschueren, K., & Colpin, H. (2023). Teacher–student relationships as protective factors in marginalized youth. *School Psychology International, 44*(2), 123–140.

Unaka, N., Hartman, M. E., & Berman, S. (2024). Addressing child health inequities through multi-sectoral policy responses. *Pediatrics, 153*(2), e2023067890.

van Dijk, M., & van Geert, P. (2022). Dynamic system approaches to language acquisition. In *International Encyclopedia of Education* (4th ed.).

Vargas, E., Sanchez, G. R., & Valdez, C. (2020). Neighborhood disadvantage and cognitive development. *Child Development, 91*(4), 1154–1172.

Wang, L., Xian, Y., Dill, S. E., et al. (2022). Parenting style and the cognitive development of preschool-aged children: Evidence from rural China. *Journal of Experimental Child Psychology*.

Wang, Y., Xie, H., & Lei, L. (2024). Peer victimization and psychological outcomes in marginalized youth. *Journal of Youth and Adolescence, 53*(1), 22–38.

Williams, C. (2021). Culturally responsive practices in special education. *Teaching Exceptional Children, 53*(4), 251–259.

Woods, K., Hashimoto, B., & Brown, E. K. (2023). A multi-measure approach for lexical diversity in writing assessments. *Assessing Writing, 56*, 100708.

Wu, S. Y., Huang, R. J., & Tsai, I. F. (2019). The applicability of D, MTLD, and MATTR in Mandarin-speaking children. *Journal of Communication Disorders*.

Yang, C., Bear, G. G., & May, H. (2023). Peer support and school belonging among marginalized adolescents. *Journal of School Psychology, 91*, 45–58.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*.

Yu, L., Huang, J., Liu, P. D., et al. (2024). How parenting styles affect the development of language skills and reading comprehension in primary school students. *British Journal of Educational Psychology*.

Zhang, D., Conger, R. D., & Sun, Y. (2022). Family resilience and child well-being in low-income households. *Journal of Family Psychology, 36*(2), 214–225.

Zhao, Y., Li, Y., & Chen, H. (2025). Multidimensional poverty and child outcomes: Evidence from longitudinal data. *Development and Psychopathology*.

Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MCVISekolahRendahwithRank.csv
- MCVIKolejVokasionalwithRank.csv
- MCVIPraUniversitiwithRank.csv
- EthicsandOthers.docx