英文文本詞彙多樣性自動化指標發展與探討

徐詩媛[1]　白鎧誌[2]　郭伯臣[2]　邱毓芳[3]
[1]亞洲大學資訊工程學系碩士在職專班
claudiahsu929@gmail.com
[2]國立臺中教育大學教育測驗統計研究所
minbai0926@gmail.com
kbc@mail.ntc.edu.tw
[3]國立臺中教育大學教師教育研究中心
yvonnechiou65@gmail.com

摘要

本研究利用 Type-Token Ratio(TTR)及 Measure of Textual Lexical Diversity(MTLD)建置詞彙多樣性自動化指標並探討台灣國中英語教科書之詞彙多樣性與各年級之間之關聯度。本研究使用國中 7~9 年級英語文本進行文本詞彙多樣性指標分析，其研究結果如下：

1. TTR 值易受文本長度的影響。當文本長度越長，文本詞彙(token)數量增加，但詞彙類型(type)數量無法如同 token 般無限制的增加，造成 TTR 值因文本長度增加而逐漸下降，因此 TTR 值只能分析較短文本，無法分析較長文本。

2. 將文本長度控制後(35 字以內)，TTR 值即隨著年級增加而上升，由此可見，計算 TTR 值時，要有文本長度的限制。在限制文本長度後，本研究發現詞彙多樣性大致上與年級有一致性的趨勢。

3. 為避免文本長度影響 TTR 值而改良的詞彙多樣性指標 MTLD，的確也得出數值趨勢也隨著年級增加而上升，因此可知，MTLD 的計算不會受到文本長度的影響。

4. 由上述三點可得知，詞彙多樣性指標數值大致上與年級高低有一致性。唯九年級(第五冊和第六冊)可能受高中職免試升學及國中基測即將在九年級下學期初舉行的影響，課程重點放在前四冊，而顯得數值稍微下降。但九年級下(第六冊) 限制文本長度的 TTR 值和 MTLD 值仍然較高，顯示詞彙多樣性指數與年級高低有一致性。

關鍵詞：詞彙多樣性、自動化指標建置、Type-Token Ratio、Measure of Textual Lexical Diversity。

Abstract

In this study, Type-Token Ratio (TTR) and the Measure of Textual Lexical Diversity (MTLD) develop lexical diversity automated text indicators and examine the lexical diversity relevance in each book of Junior English textbook.
The results are as follows:

1. The indices of lexical diversity are sensitive to the text length. TTR value was decreased as the length of text increasing.

2. After controlling for the length of the text (35 words or less), TTR value was increased with grade rising. In the text length limit, the study found that lexical diversity broadly consistent with the trend in grade.

3. MTLD was found to be the least affected by text length; therefore, it can be used for analyzing textual diversity in various length of the texts.

Keywords: Indicators of Lexical Diversity; Type-Token Ratio ; Measure of Textual Lexical Diversity

1. 前言

英語為國際語言，在學習英語的過程中，許多學習者藉由老師或師長的鼓勵，養成良好的閱讀習慣，成為高成就習得者[1]。Timothy Bell(1998)提出英文閱讀習慣對 EFL 學生的幫助，包括能增進外語理解力、增加外語習得常識、養成對字彙、文法的敏銳直覺，引起學生對外語學習的興趣及獲得延伸資訊，並能快速習得新知等優點[2]。因此，除了教師們多鼓勵學生閱讀，如何有效率且精準的為英語學習者選擇適當的文本，已成為教師們所重視的課題。以閱讀動機來看，只有難度適中的文本才能使學生產生閱讀動機[3]。

影響文章的難易度有許多因素，如文本字數、平均字長、平均句長、詞頻等表面的語言特徵；也有其他多面向的特徵，例如：文章凝聚力、文本的連貫性、讀者對文本的先備知識、詞彙多樣性、潛在語意分析等[4]。其中詞彙多樣性是指文本中不同字彙被使用的程度，當所計算出來的數值愈高代表文本詞彙有較高的多樣性，也就是文本難度較高[5]。

詞彙多樣性應用在不同的領域，從文體學、神經病理學到第二外語的學習。以語言學習來說，在國外有許多研究者利用詞彙多樣性來量化文本做不同領域的研究，例如 Daller 等人(2003)[6]研究雙語環境詞彙多樣性的特性、Carrel 和 Monroe(1993)[7]以詞彙多樣性來評估學生學習風格、Ransdell 和

Wengelin(2003)利用詞彙多樣性來預測兒童閱讀與寫作能力[8]。

在語文領域中，國外已有多位研究者利用詞彙多樣性指標來測量說話者或寫作者的詞彙運用的程度[9][10][11][12]，因此詞彙多樣性一直是語言學研究文章難易度的指標之一。國內也有研究者以詞彙多樣性作為中文文本分析指標[13]，因此本研究將以詞彙多樣性為主來探討國中英文文本。

目前國內較少以英語作為第二語言來分析文本的研究，因此本研究目的為建置自動化分析指標可以提供英語文本的訊息給教師作為閱讀教學之依據。本研究目的如下：

I. 建置詞彙多樣性自動化文本分析指標。

II. 探討詞彙多樣性與國中英語教科書冊別之關聯性。

2. 文獻探討

2.1 Coh-Metrix

Coh-Metrix 是由曼菲斯大學所發展一套有別於傳統可讀性公式只考慮文中詞彙或句子理解等表面特性的多樣化文本電腦自動化分析系統。自 2002 年開始發展，目前已開發至 Coh-Metix 3.0 版[14]。在 3.0 版中，提供更多提供分析文本的指標，總計 11 個向度，106 個指標。表 1 為 Coh-Metrix 3.0 版指標內容。

表 1. Coh-Metrix 3.0 版指標

	向度名稱	指標數量
1	描述性 Descriptive	11
2	文本易讀性分數 Text easability Principal Component Scores	16
3	參照凝聚力 Referential Cohesion	10
4	潛在語意分析 LSA	8
5	詞彙多樣性 Lexical Diversity	4
6	關聯詞 Connectives	9
7	情境模式 Situation Model	8
8	句型困難度 Syntactic Complexity	7
9	句型密度 Syntactic Pattern Desity	8
10	詞彙訊息 Word Information	22
11	可讀性 Readability	3

這些指標可用在不同的方式來測量文章的凝聚力(cohesion)與連貫性(coherence)。所謂凝聚力(cohesion)指文本的特性，由文本中明確的文字、片語、句子來幫助讀者理解文本意義。連貫性(coherence)指讀者對文本內容所產生的心裡表徵。Coh-Metrx 則是提供文本凝聚力(cohesion)的研究指標[14]。

本研究主要是針對 11 個向度中的詞彙多樣性進行文本分析指標建置並以詞彙多樣性來探討國中教科書各冊別間的關聯性。

2.2 詞彙多樣性

詞彙多樣性指一個文本中所使用的詞彙的豐富性。當文本以只標算出的數值越高，表示所用詞彙類型較多、重複性少，具要較高的詞彙多樣性。在 Coh-Metrix 中，詞彙多樣性分別以實詞 TTR、所有詞 TTR 及 MTLD 計算數值，當數值越高，代表詞彙重複越少，也就是文本難度較難 [4]。

研究詞彙多樣性已有很長一段時間，Skinner(1937)[15]就以使用頻率製作詞彙表，並以此表建置詞彙關聯性。Carrol(1938)[16]在一年後研發出第一個測量詞彙多樣性的數學方法，*k*；*k* 是利用使用詞彙總數與詞彙不同類型數的關係發展而成，用來測量學生寫作的詞彙多樣性。從此，詞彙多樣性就成為一大研究課題，而一個可靠且精準的測量方式可以引導出許多重要的資訊。[8]

而目前量化文本的詞彙多樣性最大的困難就是文本的長度。自 1940 年起，TTR 就是最常用的詞彙多樣性量化方式，但隨著文本長度的增長，詞彙出現的次數和機會就會越大，造成 TTR 值降低；如此一來將會影響數值的計算，以致於無法達到良好的數據結果。因此許多研究者也針對這樣的問題而發展出不同的方法來計算文本的詞彙多樣性[8]。除了傳統的 TTR 之外，還有由 McKee 等人(2000)利用 vocd 的 D 數值來計算文本的詞彙多樣性，D 的值越大，代表該文本的詞彙多樣性越高[17]；而由美國曼菲斯大學發展的 Coh-Metrix 電腦自動化分析系統的 106 個指標中的 MTLD 即是針對 TTR 易受文本長度影響而採用的量化指標。

2.3 TTR (Type-Token Ratio)

TTR 做為測量詞彙多樣性的傳統方法。type 指文本中不重複詞彙的類型；例如"The old man loves the young man." 中，type 即指 the, old, man, loves, young，共五種類型。token 指該文本中總詞彙數量，以例句來說，token 即指 the, old, man, loves, the, young, man，共七個詞彙。

TTR 的歷史要推到 1940 年代，當時正發展口說及寫作的量化指標[15]。但時至今日，研究者發

現若不統一文章長度，TTR 的數值將無法正確比較文本間的詞彙多樣性[8]。因此，除了利用傳統 TTR 測量文本詞彙多樣性外，配合其他計算方式才能有效測量文本詞彙多樣性。

TTR 計算方式是以文本中不同詞彙的數量除以所有的詞彙，得出的數值即該文本詞彙多樣性的程度，也就是數值越高代表多樣性的程度越高；也就是使用詞彙較多樣、豐富； 當文本詞彙多樣豐富，即可代表該文本難度較高。

2.3 MTLD (Measure of Textual Lexical Diversity)

詞彙多樣性的計算最大問題在於文本長度。當文本長度越長，文本總詞彙量(token)增加，但與詞彙類型(type)的比值卻降低，導致文章越長，TTR 值越低，就代表文本難度越低；這樣的結果無法正確計算出該文本的詞彙多樣性。

為解決文本長度造成 TTR 數值降低而誤判的現象，有研究者發展 MSTTR 的方法來計算詞彙多樣性。MSTTR 的計算方式是以相等詞彙數分割文本，然而若以不同詞彙數分割同一文章會計算出不同的 MSTTR 值[8]，因而降低 MSTTR 的信度，因此以字串為單位來做運算的 MTLD 就成為計算詞彙多樣性的指標之一。

MTLD 並沒有一定長度的字串，而是逐字計算每個字的 TTR 值。當依序計算到個詞彙時，TTR 值降到了 0.72，則作為文本的切割點；每切割一次，就劃分出一個字串；再由下一個詞彙繼續開始計算 TTR 值、切割文本，直到文本結束。但當此字串不足十個詞彙時，在計算 MTLD 時將不予計算。為提高 MTLD 的信度卻不會破壞文本的順序架構和文本的完整性，所有計算方式將由文本最後一個字計算並切割回到文本第一個字[5]。

3. 研究方法

本研究應用實詞詞彙和所有詞的 Type-Token Ratio(TTR)以及所有詞的 Measure of Textual Lexical Diversity(MTLD)來發展多樣性自動化文本分析指標，並使用現行國中教科書作為本研究發展的指標分析之探討，本研究發展指標為：

I. Type-Token Ratio, Content word (實詞)

II. Type-Token Ratio, All word (所有詞)

III. Type-Token Ratio, limited (限制文本長度)

IV. Measure of Textual Lexical Diversity (文本詞彙多樣性測量)

TTR 與 MTLD 指標計算方法如下於下列各小節分別敘述。

3.1 TTR 計算方法

TTR 是計算字數在文本中出現的頻率，將文本中 type 除以 token 後的比值決定字彙多樣性的程度[19]。type 的定義是指一段文本中的字彙類型數，一個字重複出現仍為一個"type"；token 定義則是指一段文本中所有的字彙數，稱為"token"。

舉例來說：若"apple"在文本中出現五次，type 值為 1；token 值為 5。當 TTR 值是 1 時，代表每字在文本中出現一次，閱讀理解相對來講比較困難。若 TTR 值減少，文本中字彙重複次數增多，即代表讀者對文本處理的容易度和速度增加，也可以稱該文本難度較低，讀者較容易理解。其 TTR 公式如下：

$$TTR = \frac{types}{tokens} \tag{1}$$

3.2 MTLD 計算方法

相較於最為人所知之測量文本詞彙多樣性的 TTR，為避免因為文本長度而造成效度不佳的狀況，MTLD 提供以連續字串來維持 0.72 的 TTR 值[12]。

步驟一：計算 MTLD 過程中，由篇首開始依序算出文本中的每一個字彙的 TTR 值，當字彙的 TTR 值達到或小於 0.72 時，切割文本，文本至此稱該字串為一個因子（factor）；下一個字再重新開始計算字彙 TTR 值，待 TTR 值達到或小於 0.72 時，再次切割文本，重新開始計算字彙 TTR 值並劃分數個因子直到文本結束。其中若一個 factor 中字彙數少於 10 個字，該串數值不使用。文本最後計算之因子並不完整，故需計算不完整因子分數以提高 MTLD 之可信度，其公式如下：

$$IFS = \frac{1 - RS}{1 - 0.72} \tag{2}$$

(IFS:不完整因子分數；RS:文本最後 TTR 值)

步驟二：將文本的總 token 數除以因子數與不完整因子分數之和，即算出第一個 MTLD1 值。

步驟三：從文本最後一個字彙依上述方式計算至第一個字，得出第二個 MTLD2 值。

步驟四：取兩個 MTLD 值的平均值，其公式如下：

$$MTLD = \frac{MTLD1 + MTLD2}{2} \tag{3}$$

3.3 文本分析

本研究所分析的文本來源為本研究建置的英語語料庫中的國中 7~9 年級文本，共有 414 篇文

章，針對英文教科書文本，計算 7~9 年級各年級文本的平均 TTR 值與 MTLD 值並進行趨勢探討。

4. 研究結果

本研究採用四家出版社 7~9 年級英語教科書文研究文本，分別計算 7~9 年級文本中詞彙的平均 TTR 值。結果如圖 1 所示，實詞的平均 TTR 值在七年級下及九年級上時數值降低，其他年級也未看出明顯的趨勢；也就是說在實詞 TTR 值並未出現隨年級增加而逐漸上升的趨勢。

表 2. 7~9 年級英語文本實詞 TTR 值

	TTR 值					
	七年級上	七年級下	八年級上	八年級下	九年級上	九年級下
H	0.822	0.794	0.794	0.833	0.745	0.797
N	0.852	0.796	0.805	0.817	0.779	0.778
K	0.830	0.773	0.799	0.857	0.776	0.746
L	0.771	0.755	0.839	0.806	0.793	0.739

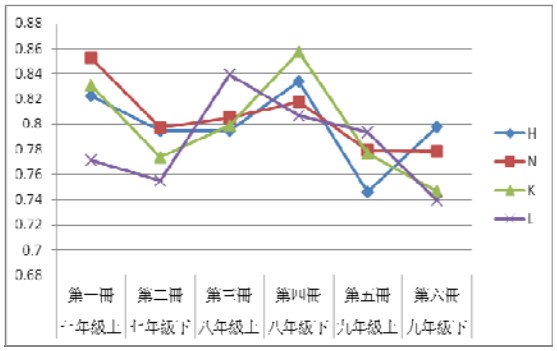

圖 1. 7~9 年級實詞 TTR 趨勢圖

表 3 是分別計算文本中所有詞的平均 TTR 值，由圖 3 的趨勢圖顯示結果與圖 1 類似，所有詞的平均 TTR 值並未隨著年級的增加而逐漸上升。與表 2 比對來看，所有詞的 TTR 值要比實詞 TTR 值還要低。

表 3. 7~9 年級英語文本所有詞 TTR 值

	TTR 值					
	七年級上	七年級下	八年級上	八年級下	九年級上	九年級下
H	0.583	0.612	0.655	0.635	0.564	0.604
N	0.623	0.589	0.659	0.641	0.560	0.579
K	0.650	0.609	0.624	0.684	0.623	0.600
L	0.598	0.613	0.652	0.621	0.607	0.577

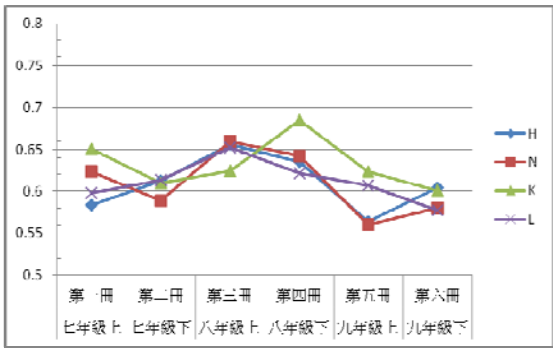

圖 2. 7~9 年級所有詞 TTR 趨勢圖

表 4 是從語料庫中挑選出文本長度在 35 字以內的文本為樣本，計算所有詞的平均 TTR 值。圖 3 結果發現限制文本後的所有詞 TTR 值會隨著年級的增加而逐步增加。由此可得知，當限制文本長度時，TTR 值有隨年級增加而數值增加的趨勢；如此可得知，文本長度確實會影響 TTR 值。

表 4. 7~9 年級英語文本所有詞 TTR 值(限制文本長度)

	TTR_limited 值					
	七年級上	七年級下	八年級上	八年級下	九年級上	九年級下
H	0.722	0.764	0.826	0.846	0.830	0.845
N	0.698	0.758	0.830	0.834	0.814	0.823
K	0.764	0.767	0.807	0.819	0.808	0.821
L	0.682	0.761	0.828	0.833	0.831	0.826

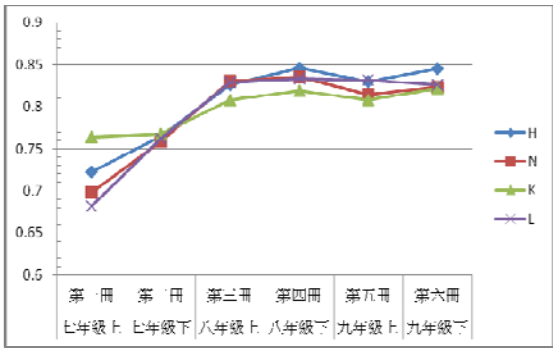

圖 3. 7~9 年級英語文本所有詞 TTR 值趨勢圖(限制文本長度)

表 5 是利用 MTLD 計算文本詞彙多樣性與年級之間的關係，根據圖 4 所發現，隨著年級增加，MTLD 的值確有增加的趨勢。如圖 3 類似，限制文本的所有詞 TTR 值和 MTLD 值都在八年級下，也就是第四冊時到達最高值，推斷應該是為配合九年級下將進行高中職免試升學以及國民中學基本學力測驗的舉行，而將國中英語課程重點著重於第四冊，如此有助於九年級能有足夠時間申請免試入學及複習國中課程。

表 5. 7~9 年級英語文本 MTLD 值

MTLD 值						
	七年級上	七年級下	八年級上	八年級下	九年級上	九年級下
H	37.12	47.20	70.94	72.21	76.74	80.87
N	38.55	44.59	70.73	81.56	77.49	84.55
K	45.06	52.51	66.55	89.93	72.91	74.89
L	33.49	46.78	74.05	73.62	77.88	74.32

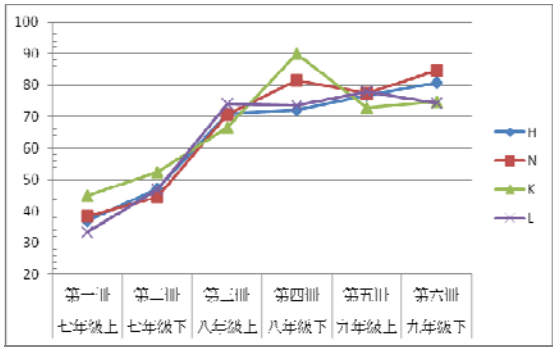

圖 4. 7~9 年級 MTLD 趨勢圖

5. 結論與未來建議

本研究欲建置英文文本詞彙多樣性自動化指標並探討詞彙多樣性與國中英語教科書中各冊別之間關聯性，經資料分析結果後本研究針對以下幾點結論說明。

5.1 TTR 值會受文本長度影響

本研究結果發現，以 TTR 計算詞彙多樣性時，不論只計算實詞的多樣性或是所有詞的多樣性，TTR 值在七年級上(第一冊)時數值反而較高；其他年級，尤其是九年級上(第五冊)及九年級下(第六冊)時，數值較低。可推論 TTR 值會隨著文本長度增加而逐漸下滑。若將文本長度限制後(35 字以下)，TTR 值出現隨著年級增加也逐漸增加的趨勢，與過去的研究結果呈現一致性[12][9]。因此，TTR 運算值的確受到文本長度的影響，只能選擇運用在較短文本或限制字數的分析。

5.2 詞彙多樣性與年級間有一致性的趨勢

本研究發現，不論是由限制文本長度的 TTR 值或 MTLD 的計算結果，所得到的值大部分會隨著年級增加而有逐漸上升的趨勢，尤其以七年級上(第一冊)到八年級下(第四冊)上升幅度最為明顯，顯示詞彙多樣性與年級之間有一致性的趨勢。也就是說，越高年級，文本生字量增加，以致於詞彙重複機會低，造成文本難度增加。

至於九年級上(第五冊)的限制文本長度的 TTR 值和 MTLD 值相較於八年級下(第四冊)的 TTR 值與 MTLD 值，並未如預期上升，反而稍微下降，應該是受到高中職免試入學申請與國中基本學力測驗即將在九年級下的學期一開始就進行的影響。為配合九年級下的高中職免試入學申請以及國中基測，九年級課程以複習為主，推測各版本出版商有相同的默契將課程主軸著重於升上九年級(第五冊和第六冊)之前。另外，由圖 3 及圖 4 的趨勢圖也可以看出，九年級上(第五冊)與九年級下(第六冊)的曲線符合年級增加，限制文本長度的 TTR 值和 MTLD 值也隨之增加的一致性。

5.3 未來研究建議

本研究主要是建置英文文本詞彙多樣性自動化指標，經教科書文本分析結果發現年級愈高，其詞彙多樣性程度會越複雜，未來可以納入國中英文課外閱讀文本應用 TTR 與 MTLD 自動化指標分析，可以進一步驗證國內英文文本的詞彙多樣性程度，作為英文教師在實施閱讀教學讀本篩選之參考。

參考文獻

[1] 廖悅淑。EFL(English as a Foreign Language)低成就學習這如何培養英閱讀習慣。基隆市閱讀教育電子報，第六期。

[2] T. Bell, “Extensive Reading: Why and How? ”, The Internet TESL Journal, Vol. 4, No. 12, http://iteslj.org/Articles/Bell-Reading.html, 1998.

[3] J. S., Chall and S. S. Conard, “Should textbooks challenge students? the case for easier or harder textbooks,” New York: Teachers College Press, 1991.

[4] A. C. Graesser, D. S. McNamara, , M. M. Lauwerse and Z. Cai, “Coh-Metrix: Analysis of text on cohesion and language,” Behavorial Reserch Methods, Instruments, and Computers, 36, 2004, pp. 193-202.

[5] P. M. McCarthy and S. Jarvis, “MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment,” Behavior Research Methods, 42, 2010, pp. 381-392.

[6] H. Dalle, R. Van Hout, & J. Treffers-Daller, “Lexical Richness in the spontaneous speech of Bilinguals,” Journal of Applied Linguistics, 3, 2003. pp.197-222.

[7] P. L. Carrell, & L. B. Monroe. “Learning styles and composition,” The modern Language Journal, 77, 1993. pp.148-162.

[8] P. M. McCarthy, “An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD) ,” PhD Dissertation, The University of Memphis, 2005.

[9] T. Santos., “Professors' Reactions to the Academic Writing of Nonnative-Speaking Students,” TESOL Quarterly, Vol.22, No.1, 1998, pp. 69-90.

[10] B. Laufer, and P. Nation, “Vocabulary size and use: Lexical richness in L2 written production,” Applied Linguistics, 1995, pp. 307-322.

[11] C. A. Engber., “The relationship of lexical proficiency to the quality of ESL compositions,” Journal of Second Language Writing, 1995, pp. 19-155.

[12] D. D. Malvern, B. J. Richards, N. Chipere and P. Duran, “Lexical Diversity and Language Development,” Quantification and Assessment. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan, 2004.

[13] 葉靜如，中文文本詞彙多樣性自動化分析系統建置與探討，IETAC 2013 資訊教育與科技應用研討會，2013。

[14] Coh-Metrix, http://cohmetrix. memphis. edu/cohmetrixpr/index.html.

[15] B. F. Skinner. “The distribution of associated words. Psychological Record,” 1, 1937, pp.71-76.

[16] J. B. Carroll. “Diversity of vocabulary and the harmonic series law of word-frequency distribution,” Psychological Record, 2, 1938, pp. 379-86.

[17] G. McKee, D. Malvern, & B. Richards. “Measuring vocabulary diversity using dedicatd software,” Literary and Linguistic Computing, 15, 2000, pp. 323-37.

[18] W. Johnson. “Studies in language behavior: I. A program of research,” Psychological Monographs, 56,1944, pp.1-15.

[19] R. Koizumi, “Relationships between text length and lexical diveristy measures: Can we use short texts of less than 100 tokens?,” Vocabulary Learning and Instrucion, 1(1), 2010, pp. 60-69. doi: 10.7820/vli.v01.1.koizumi.