月例研究報告

Type-Token Indices and Lexical Diversity

Dax Thomas

Lexical diversity is the degree to which individual words are repeated (or not repeated) in a given text. Measures of lexical diversity can be used by EFL teachers in the selection of graded readers and content textbooks for their classes, or for assisting in the evaluation of active vocabulary usage in student written and spoken discourse. They can also be used by literary critics in estimating a particular author's overall vocabulary size, and can play a role in forensic linguistics and authorship identification.

More specifically, lexical diversity in a text is the relationship between the number of tokens (words) and the number of types (types of word) in that text. Traditionally, this relationship was presented in the form of a simple type–token ratio (TTR): the total number of types divided by the total number of tokens in the text. These raw type–token ratios, however, are problematic in that they are highly dependent on the text's length making it difficult to accurately compare texts of different sizes. In order to better compare texts of differing length a normalization tool is needed. Several TTR normalization attempts have been made using both mathematical approaches (Root TTR, Corrected TTR, Log TTR; Malvern et al., 2004), and random sampling approaches (voc–D, HD–D, MTLD; McCarthy and Jarvis, 2010), but these fall short of solving the problem satisfactorily.

This report briefly introduces an alternative normalization tool, the Type–Token Reference Curve (TTRC), and explores two ways in which it might be utilized: the Type–Token Reference Curve Index (TTRCI) and the Type–Token Area Index (TTAI).

Type–Token Curves

A basic Type–Token Curve (TTC; Youmans, 1990) is a graph representing the lexical diversity of a text where the running type–count is plotted against the running token–count (Fig. 1).

A steep curve represents higher lexical diversity while a shallow curve represents more vocabulary recycling. This "bird's eye view" of the lexical diversity of a text can be quite useful when studying a small number of texts of similar length.   However, if the texts are of greatly differing lengths (Fig. 2) or there are a large number of texts to be compared (Fig. 3) then using TTCs alone becomes problematic. To help with this problem a Type–Token Reference Curve can be used.

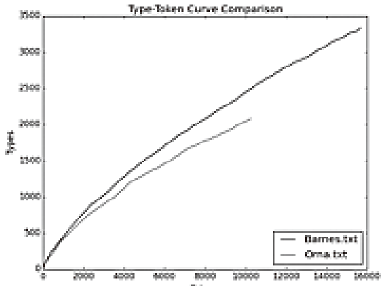

Fig. 1: Type–Token Curves for two content English textbooks

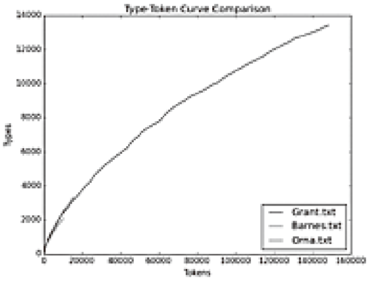

Fig. 2: Comparing long and short texts.

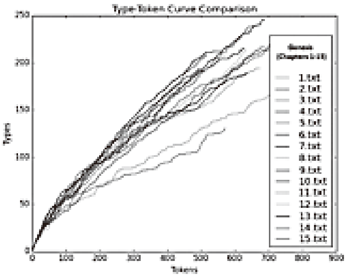

Fig. 3: Comparing a large number of texts.

Type-Token Reference Curve

A normative 300,000-token Type-Token Reference Curve (Fig. 4) was constructed using the average type counts of 10 long public domain books. This curve represents the normal tendency for the introduction of new vocabulary to decrease as text length increases.

Type-Token Reference Curve Index

The TTRC can then be used to generate a Type-Token-Reference Curve Index. The type count at the end of the text is divided by the type count at a point on the reference curve matching the token count of the text in question (Fig 5).

Indices above 1.000 indicate a lexical diversity above the norm, while those below indicate a text with less than typical lexical diversity. These indices allow for convenient comparison of a large number of texts (Table 1).

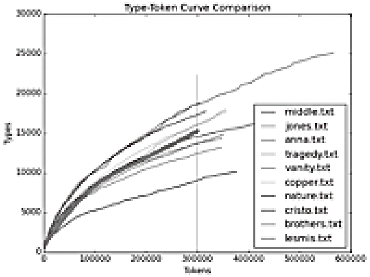

Fig. 4: Normative curve created by averaging type counts of 10 different books.

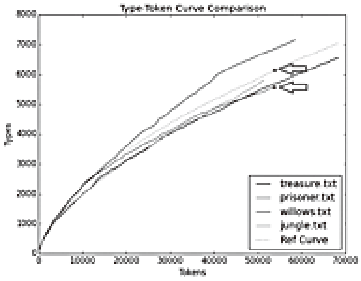

Fig. 5: Points used for calculating TTRCI.

Table 1: TTRC Indices of 4 novels.

Text	TTRCI
Wind in the Willows.txt	1.109
The Jungle Book.txt	0.975
Treasure Island.txt	0.929
Prisoner of Zenda.txt	0.909

Type-Token Area Index

Another issue worth considering in lexical diversity studies is the rate at which new vocabulary is introduced. Changes in content throughout the length of a text may result in parts of the text having different lexical diversity. For example, the TTC for *Treasure Island* begins shallower than that of *Prisoner of Zenda* but crosses over it closer to the end of the book (Fig 5). In order to better represent this variation in the introduction of vocabulary throughout the length of the text, the same reference curve can be used to calculate a Type-Token Area Index. The area under the curve of the text (Fig. 6) is divided by the area under the reference curve (Fig. 7) at a point matching the token count of the text in question to arrive at the TTAI (Table 2).

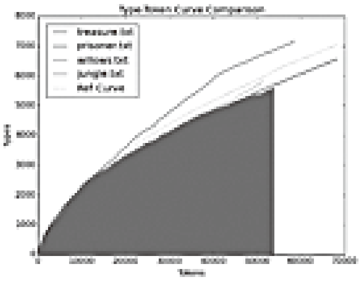

Fig. 6: Area under the curve for text.

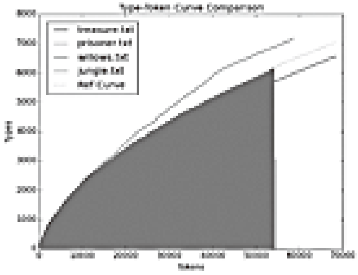

Fig.7: Area under the curve for reference curve.

Table 2: TTA Indices of 4 novels.

Text	TTAI
Wind in the Willows.txt	1.097
Prisoner of Zenda.txt	0.948
The Jungle Book.txt	0.930
Treasure Island.txt	0.920

When looking at TTRCI, *Treasure Island* has a higher index (and therefore a higher lexical diversity) than *Prisoner of Zenda.* However, when looking at TTAI, *Prisoner of Zenda* is shown to be much more lexically diverse. This correlates much better with the TTC bird's eye view showing *Prisoner of Zenda* to have a steeper curve for roughly 90% of the length of the text.

References

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

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.

Youmans, G. (1990). Measuring lexical style and competence: the type–token vocabulary curve. Style, 24 (Winter), 584–599.