Measures of Lexical Richness

語彙統計の基礎

投野由紀夫（東京外国語大学）

VOCABULARY KNOWLEDGE

Lexical richness = " how many different words are used in a text (spoken or written)."

Lexical diversity

- The proportion of individual words in a text
- TTR is the most typical LD measure.

Lexical density

- The proportion of lexical words in the whole text

Lexical sophistication

- The proportion of advanced words in a text

代表的な語彙統計指標

Name (author)	Year	Formula	Notes

Type token ration
= TTR (Templin)
	1957	$TTR\ (N) = \dfrac{V\ (N)}{N}$	N = number of tokens
			V = number of types

Mean word frequency = MWF (Tweedie & Baayen)	1998	$MWF\ (N) = \dfrac{N}{V\ (N)}$	N = number of tokens
			V = number of types

Lexical richnessの測定の最も基本的な考え方は，テキスト中にどのくらい異なる単語が入っているか，ということである。それがTTRの表す中心的意味である。しかし，TTRはテキスト長に依存する。

そこで，単純に異なり語ごとに平均何語現れたかを見る，Mean Word Frequency (MWF) という指標も提案されている。しかし，テキストが大きくなればなるほど単語の出現率は低くなるのは解消できない。

TTR and its variants

Name (author)	Year	Formula	Notes

R (Guiraud)	1954	$R = \frac{V\ (N)}{\sqrt{N}}$	N = number of tokens
			V = number of types

C (Herdan)	1960	$C = \frac{\log V\ (N)}{\log N}$	
a^2 (Maas)	1972	$a^2 = \frac{\log N\ - \log V (N)}{log^2\ N}$	Modification of k
Uber U (Dugast)	1978, 79	$U = \frac{log^2\ N}{\log N\ - \log V\ (N)}$	Notational variant of k

Vocd measures

Name (author)	Year	Formula	Notes
D (Malvern & Richards)	1997	TTR = (2/DN) [(1+DN)1/2 – 1]	TTR の曲線に適合するように D の値を決めるcurve fitting approach.
Vocd-D (McKee et al.)	2000	Vocd 専用のソフトを使用	Dを100回ランダムサンプル（35-50 tokens）から出し，Dの平均を出す。それを3回繰り返す。最後のスコアは10から100の間で，スコアが高いほどdiversityがある。

MTLD (McCarthy & Jarvis 2010)

MTLD = the measure of textual lexical diversity

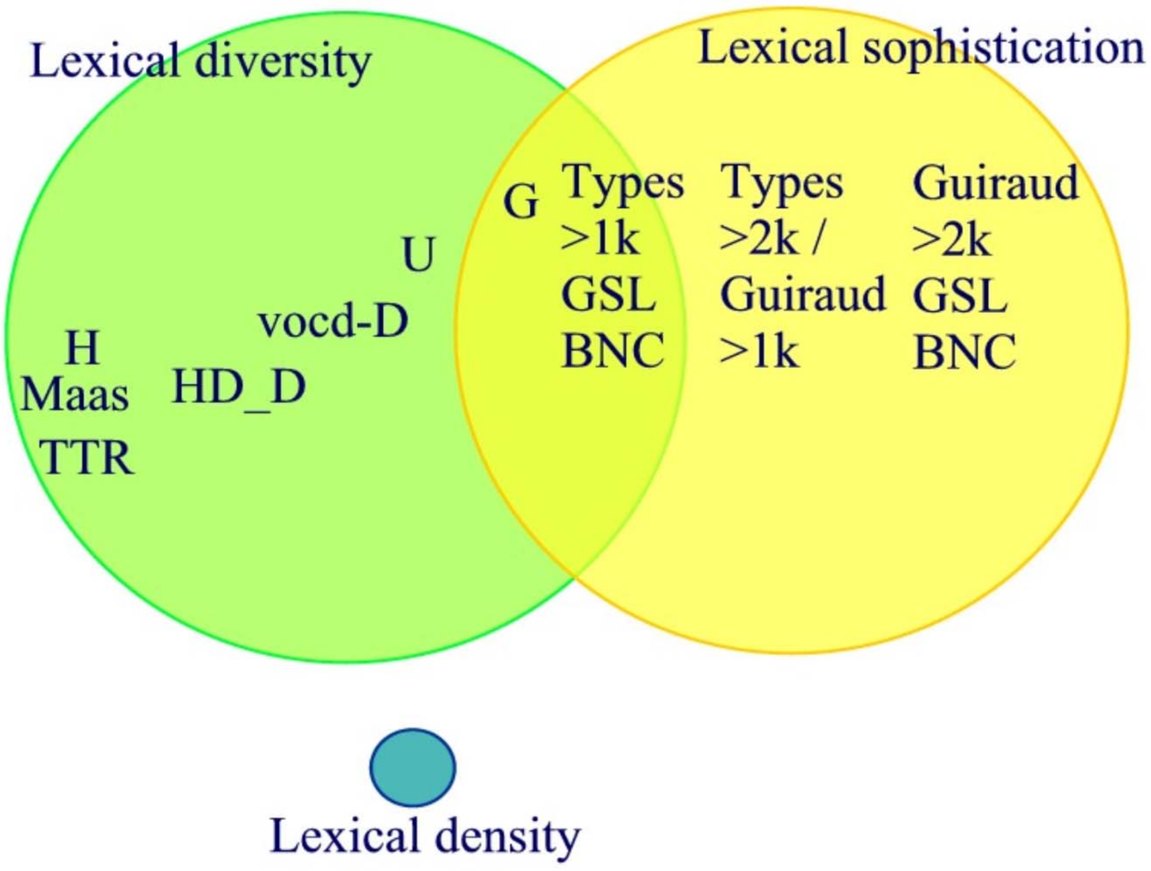
1) Cut the texts into sequences which have the same TTR (set to 0.72).

2) Calculate the mean length of the sequences which have the given TTR.

3) If the score is higher, the text is more diversified in terms of vocabulary.

The MTLD does not depend on text length in the 100 – 2,000 word range.

Relationships between the different measures



Zdislava Šišková (2012)
Lexical richness in EFL students' Narratives. *Language Studies Working Papers* 4, 26-36. University of Reading.

Significant types of lexical measures (Wolfe-Quintero, Inagaki, and Kim 1998)

- **Accuracy measures:**
- The number of error-free T-units
- Error-free T-units per T-unit
- The number of errors per T-unit

- **Complexity measures:**
- The number of clauses per T-unit
- The number of dependent clauses per T-unit
- The number of dependent clauses per total number of clauses

- **Lexical variation:**
- Type-token ratio
- Guiraud index
- D-value
- Sophisticated word type ratio
- Lexical Frequency Profile