Vocabulary Richness: Word Type/Token Ratio or Yule's K

A way of assessing the vocabulary difficulty of texts is to measure their lexical richness. Two indices one could use are Yule's K or Yule's I.

These two measures are defined as follows:

1. Yule's K = 10,000 * (M2−M1) ÷ (M1 * M1)
2. (2) Yule's I = (M1 * M1) ÷ (M2−M1) where M1 is the number of all word forms in a and M2 is the sum of the products of each observed frequency to the power of two and the number of word types observed with that frequency (Oakes 1998:204). For example, if one word occurs three times and four words occur five times, M2 = $(1*3^2) + (4*5^2)$ = 109.

The larger Yule's K, the smaller the diversity of the vocabulary (and thus, arguably, the easier the text). Since Yule's I is based on the reciprocal of Yule's K, the larger Yule's I, the larger the diversity of the vocabulary (and thus, arguably, the more difficult the text).

References

Michael P. Oakes. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

C.U. Yule. 1944. *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.