Appendix 2--Vocabulary Analysis[1]

Story	Number of Words in Analyzed Text[2]	% of Words					Average Tokens/ Family (on-list only)[6]	# Proper Nouns & Non-English Items (Fam/Tokens)
		K-1 First 1000 Word Families[3,4]	K-2 2nd 1000 Families	K3 thru K10	K11 thru K20	Off-list[5] Words		
Sugihara	859	88.0	5.4	6.5	0.0	0.1	3.1	32/80
Yunus	1705	90.2	7.3	2.3	0.0	0.1	4.2	19/60
Richard	1349	84.1	9.6	5.9	0.1	0.3	3.2	26/110
Mabaso	957	86.7	6.7	5.7	0.2	0.6	3.1	22/71
Santos	1406	82.0	7.3	9.7	0.5	0.4	3.5	11/47
Maathai	1346	86.6	9.2	4.0	0.1	0.2	3.4	8/54
Jantraka	1326	88.3	7.6	3.8	0.3	0.0	3.2	32/86
Khurana	1462	86.9	7.0	5.7	0.3	0.2	3.3	18/53
Tse	1965	83.9	10.1	5.8	0.1	0.1	3.9	30/93
Rosa	2108	85.4	7.5	6.6	0.2	0.3	4.0	14/66
Tenberken	1669	85.6	8.9	4.9	0.3	0.3	3.4	26/101
Roy	2175	84.8	7.4	9.3	1.2	0.3	3.9	24/53
Masih	1739	88.1	6.5	4.7	0.1	0.6	3.9	20/88
Guerrero	1854	82.4	9.1	7.5	0.2	0.7	3.6	25/93
Last Story								
All Stories together (not an average)	21915	85.8	8.0	5.7	0.3	0.3	11.5	310/1058

For more information, see Cobb, T. *Why and how to use frequency lists to learn words*. http://www.lextutor.ca/research/

---

[1] Submitted to http://www.lextutor.ca/vp You can submit any story there for analysis of vocabulary.

[2] Text submitted for analysis excluded picture legends, textboxes, questions, references, many proper nouns, and non-English words and expressions (for example http, www and other website name components, pdf, names of countries).

[3] K1=1000 most-frequently occurring word families in text database, K2=2nd most-frequently occurring thousand, K3=3rd most-frequently occurring thousand, etc.

[4] Family = group of words from same stem (eg attend, attendee, attending, attendant, attendance); Token = individual word in text Tokens/family (tokens per family) is given as a measure of the repetition in the text (note however that it is an average figure.)

[5] Off-list = not on the list of words compiled from the text database and includes some currently very common words such as internet, website, and email (reflecting the fact that the database was compiled some years ago). Hyphenated words tend to end up here also since the analysis strips out the hyphens and then doesn't recognize the resulting word.

[6] Token = individual word