CALIFORNIA STATE SCIENCE FAIR
2009 PROJECT SUMMARY

Name(s)
Ishan S. Puri

Project Number
S0312

Project Title
Lexical Distributions and Electronic Literacy: A Corpus Linguistic Analysis of Textual Richness

Abstract

Objectives/Goals
Little quantitative work has been published on electronic-mediated communication and electronic literacy. This paper hopes to identify the textual richness of representative corpora that will lead to conclusions about the literacy as a whole. My goal is to quantitatively analyze this literacy through the creation of programs written in Python. Regressed Zipfian coefficients, vocabulary size, 4-grams, and hapax legomenon will be analyzed to embody textual richness. With my results, conclusions can be reached regarding the future of language, and the debilitating or regenerative effect of electronic communications.

Methods/Materials
The Python software, SciTE, Microsoft Word, and a personal computer (2GB RAM, 2 GHz) were used. Each corpus was collected from representative, publicly-available online sources and was kept at 50,000 words, in total a collection of a 250,000 words (Blogs-Global Top 100 Technorati, IM-ICQ, Email-Enron corpus, Spoken-Michigan Corpus of Academic Spoken English, Volumes-Harvard Classics). Three Python programs were created: 1. First-order Zipfian coefficients through Levenberg-Marquardt algorithm 2. Vocabulary size and 4-gram distributions 3. Hapax legomenon count. Each corpus was run through the programs in IDLE.

Results
As predicted, in order of lexical richness from greatest to least: Volume, Email, Blog, IM, and Spoken corpus. For H1, Zipfian coefficients, we find -0.91 (Volume), -1.00 (Email),-1.00 (Blog), -1.06 (IM), and -1.19 (Spoken). Lexical richness results follow directly from this data. For H2,vocabulary size: 8,958 (Volume), 8,271 (Email), 7,732 (Blog), 6,712 (IM), and 4,631 (Spoken). The 4-gram results illustrate the standardization of phrases in electronic literacy: 313 (Volumes) and 1,308 (Spoken), and in the electronic: 1,957 (Email), 1,760 (Blog), and 3,204 (IM). Finally we note the hapax legomenon in H3: 5,437 (Volumes), 4,491 (Email), 4,340 (Blog), 3,421 (IM), and 2,377 (Spoken).

Conclusions/Discussion
We conclude that electronic literacy is less rich and perhaps slightly debilitating in terms of vocabulary size. Yet rank-frequency charts amend that this new literacy is not as blunt as some suggest, and that it maintains several features of traditional communication. Therefore we conclude that electronic literacy is rather a hybrid of the spoken and written word.

Summary Statement
Lexical richness of electronic and non-electronic literacy is calculated through the development of three Python programs measuring Zipfian coefficients, vocabulary size and 4-grams, and hapax legomenon.

Help Received
Dr. Stabler introduced me to register and gave me introductory material to read.