**First Year English as Big Data:**

**Analyzing written text regarding revision and use of linguistic features**

Duncan A. Buell

In collaboration with Dr. Chris Holcomb, Dr. Qiandi Liu, Dept. of English Language and Literature

USC teaches about 125 sections of First Year English every semester, and these classes generate 5000-7000 freshman student essays in draft and final versions. What can be learned about freshman student writing from the analysis of these essays?

Since 2014 we have been collecting and analyzing FYE essays, and we now have a corpus of just under 20 thousand curated essays totalling about 40 million words. We are not doing “machine grading”. But we are analyzing these to answer various questions about student writing in hopes that the answers will lead to improvements in how we teach students to write good prose.

* What are the characteristics of revision of the draft of an essay into the final version?
* What are the linguistic features used by FYE students, and how do these differ from the features that appear in academic writing, or in good magazine writing, or in informal conversation?
* Are there observable differences between native English speakers and second-language students in their writing of English prose?
* Are there observable differences between students in the humanities and students in the sciences?

We have analyzed our large corpus and have answers to some of these. We have not yet obtained answers to all these questions, but will describe what we do and how we do it.