Help Wanted: Sharing Data for Research on Reading and Writing

Should there be a large, open collection of student writing, representing the range of ability and accomplishment among American high-school and college students today? We think so, but we’d like to hear your opinions.

“We” are a group of linguists, psychologists, computer scientists, and writing-program professionals; and we believe that that a large collection of student writing, as part of a larger collection of texts and annotations, would provide an essential basis for many important kinds of research.

Our general idea is to create an open and evolving dataset of both student writing and expert writing, combined with an open and evolving collection of layers of annotation. The annotations might be linguistic (syntax, word senses, co-reference, discourse structures), editorial (mistakes, infelicities, suggested corrections), or psychological (eye tracking, EEG or MEG, reading comprehension, readers’ evaluations, etc.). Since the collection would be a large one, not all kinds of annotation would be applied to all parts of it. And of course, not all users of the data would be interested in all types of annotation.

This collection could be used to estimate how much trouble students of different kinds at different levels have with different aspects of writing. It could be used to study the effects of writers’ choices on readers’ uptake, and might thus help to create better interactive advice for writers. And in addition to these and other innovative uses, such a collection would provide a larger and more diverse basis for standard sorts of reading research.

There are many available collections of (billions of words) of expert writing, and plenty of reading researchers who are willing to share their data, and plenty of computational linguists who are willing to share their algorithms and even their programs. The piece of the puzzle that is still entirely missing is a large and diverse collection of student writing, as well as editorial annotation, commentary, and evaluation for some of it.

In this session, we’ll sketch some of the kinds of research that such a collection would facilitate, and solicit the opinions of CWPA attendees about problems and opportunities.