

Questions Worth Asking: Intersections between Writing Research and Computational Linguistics

Anne Ruggles Gere and Laura Aull

University of Michigan
Ann Arbor, MI 48104, USA

Abstract

Rather than explain research that has already been carried out, this paper describes a specific context of writing instruction and poses questions about how research on writing and computational linguistics might be brought together to address three pressing issues: the validity of Directed Self-Placement; the relationship between confidence and competence in student writing; and strategies to help English Language Learners, especially those in the category of Generation 1.5, improve their writing.

1 Introduction

In this paper we would like to explore questions that hover at the intersection of writing research and computational linguistics. First, however, we would like to explain the context of our work since we believe that context is a shaping force in any research on writing. The site where we work is the Sweetland Center for Writing at the University of Michigan, and this Center is responsible for the placement of some 4000 first-year students in several different colleges including LSA, Nursing, Kinesiology, Art and Design, and Music.

The majority of students who enroll at this university have performed well on tests and in high school. Only approximately 5% have a high school GPA under 3.3 or a SAT verbal below 530. The university also enrolls a significant (approximately 5%) number of international students who are English Language Learners,

along with an unknown number of Generation 1.5 students who are also learning to manage various aspects of academic English but who are very difficult to identify because they do not have the clear markers of international students. The Center is also responsible for oversight of all courses that meet the First Year Writing Requirement and the Upper Level Writing Requirement.

First-year students enroll in either a developmental course, Writing 100, or one of the seven different courses that satisfy the First Year Writing Requirement (FYW). For nearly a decade students decided between the two, in concert with their advisors, by participating in a form of Directed Self-Placement (DSP).

DSP, which came into use in the United States during the final years of the 20th century, puts into student hands the decision about placement in writing courses. DSP was first implemented at colleges small enough to provide significant amounts of one-on-one time between students and advisors and/or writing instructors. DSP has taken various forms, depending upon the local context, but it always asks students to assess their own abilities as writers.

In the version of DSP first implemented at our university, students answered seven questions about their reading habits and their writing practices, along with their grades and test scores. The questions used from 2000 to 2005 gave more attention to mastery, as in “I have learned the correct forms of standard written English and make few mistakes in sentence construction,

punctuation, and usage,” while the survey used from 2006 to 2008 focused more on comfort or confidence, as in “I am comfortable using standard written English, including the correct forms of grammar, punctuation, and sentence construction.”

Significantly, the number of students who received a recommendation for Writing Practicum dropped dramatically—from approximately 1000 to approximately 200—when the second form of the survey was introduced. The decisions about enrollment, of course, remained in the hands of students, and it is worth noting that the percentage of students who followed the recommendation generated by the survey increased from 15% to 35%.

2 Questions of Validity

Analysis of the DSP process that had been in place from 2000 to 2008 (under both sets of questions) showed that it had little validity (Gere et al., forthcoming). It lacked substantive validity because of the time gap between completion of the survey and course selection; it lacked structural validity because the survey questions bore little relation to the construct of writing central in the FYW courses; it lacked validity of generalizability because only a small percentage of students followed the recommendation generated by the DSP survey; it lacked external validity because there was a low correlation between students’ scores on other measures and on the DSP survey; and it lacked consequential validity because the construct of writing operating in the Writing 100 course bore little relation to that emphasized in the DSP survey. This analysis, combined with the fact that students’ perception of the importance of writing was influenced by the contrast between answering seven multiple choice questions and completing substantive tests in math, chemistry and foreign languages, led to a reconfiguring of the DSP process to include writing an essay and answering questions about that process as well as about literacy practices more generally.

Beginning in the fall of 2009, entering first-year students at our university write an evidence-based argument in response to a 3500 word publication. Essays of entering students are submitted electronically and are delivered to

individual instructors of students’ first writing class so that they become incorporated into instruction. The Writing 100 course has been redesigned so that it is aligned with the construct of writing in the DSP process and in the FYW course.

In other words, we now have a large and growing corpus of student writing, and it would be helpful to think through how we might make best use of it, with regard to questions of validity as well as other issues. The DSP essay corpus currently includes over 3500 student essays comprising over three million words, and by the end of August 2010, these numbers will double as a new cohort of students enters the University. All the texts in the initial corpus were written by incoming first-year students in response to a prompt for an evidence-based argument about a Malcom Gladwell essay that discusses the difficulty of predicting which candidates will become good teachers—or quarterbacks or financial advisors. Instructions included a recommendation to consider these features: focus or development around a clear central thesis or argument; structure or organization that elaborates on and supports the central argument; and evidence or well-chosen examples from the text to support claims.

In addition to this corpus, we have the potential to create a smaller corpus of student writing produced in first writing classes, both Writing 100 and courses that satisfy the First Year Writing Requirement, as well as personal narratives written as part of each student’s admission portfolio. In coming years, we could also collect samples of writing across the entire undergraduate experience of a subset of students. One of the questions we would like to discuss, then, centers on what decisions we should make about structuring additional corpora so as to take best advantage of the texts and materials available to us.

One clear direction for our work is to continue the investigation of validity to determine the extent to which the modifications in the DSP process and in Writing 100 enhance the validity of the placement process now in place. In particular, it would be useful to learn more about the consequential validity of the DSP process since its main result or consequence is enrollment in either Writing 100 or a course that meets

the First Year Writing Requirement. Among the possible questions to investigate are these:

- How can we best use the existing corpus and additional ones we might create to determine the extent to which the writing of students who elect Writing 100 differs from that of students who choose to enroll immediately in courses that meet the FYW requirement?
- How might we best create subgroups (and subcorpora) to understand how writers in each subgroup articulate arguments and use evidence?

The evidence-based argument is central in both contexts of first writing courses, and the construct of writing that operates in the DSP and in Writing 100 includes features of formal, purposeful, coherent, complex, audience-aware, and evidence-based writing. A variety of rhetorical choices in academic writing help writers achieve these features; for example, we know from the work of Hyland (2005) that effective writers use textual signals to pull readers along their line of argument, so one approach in our research would be to compare the writing of students who elect Writing 100 and those who do not in terms of their use of textual signs that make the terms of their arguments clear.

Given student data and surveys we have access to, we also have the capacity to create subcorpora based on student grades and scores, English nativeness, student high school types, or students' reported attributes such as confidence or writing experience. Understanding how writers in various subgroups construct arguments will help answer key questions about the validity of the current form of DSP, and we welcome discussion of how quantitative linguistics can aid in that process.

3 Questions of Confidence

Another set of questions emerges from analysis of students' responses to the DSP survey. This examination showed that there were a few "trigger" questions that influenced students' choices about which writing course to take. That is, certain questions were the ones that propelled

the greatest number of students to take or not take Writing 100. Most prominent among these were the questions dealing with the issue of confidence, as in "I am confident about my ability to comprehend unfamiliar texts."

In a subsequent survey of students who had already enrolled in either Writing 100 or a course that meets the FYW requirement, the issue of confidence became even more prominent. When asked to rank the importance of various factors in their self-placement in a writing course, "confidence in my own writing ability" was the number one factor for the great majority of students.

The next most important factor, input from an academic advisor, received less than half as many "most important" responses. This finding is significant in at least two ways, and it also raises questions that can call upon the resources of computational linguistics. One dimension of the significance of the confidence issue is that confidence is central to the theory underlying Directed Self-Placement. The literature on DSP positions confidence as the goal of a developmental course and a desired result of a FYW course is that students will develop "writing confidence." Indeed some scholars have suggested that DSP may be more a measure of confidence than of writing ability (Reynolds, 2003). The importance of confidence is magnified by the fact that confidence is frequently equated with competence in writing; it is also credited with driving out apprehension about writing, and with enhancing the authorial identity of students.

Another significant dimension of confidence, however, troubles its relationship to DSP and to writing more generally because empirical studies show that confidence in writing does not have a fixed or stable meaning. The person who expresses considerable confidence in writing essays may experience and express a lack of confidence about writing in another genre or form such as a grant proposal or lab report. The student who is a confident writer in high school may have a significant loss of confidence when faced with the writing tasks of college or the workplace. Writers who express confidence may or may not be able to produce writing that is recognized by others as "good." And those confident writers who are recognized for "good"

writing in one context may not be so recognized in other contexts.

Confidence, which is closely allied with self-efficacy, is task specific, and this complicates the meaning of writing confidence. Given the importance and instability of confidence in relation to writing and to DSP specifically, it will be useful to learn more about how confidence is manifested in student writing. Because the research on the relationship between confidence and competence in writing is mixed, it will be important to explore this relationship more closely. The corpus of student writing along with information about the questions to which students respond, particularly those focused on confidence, allows us to compare patterns in subcorpora of writing done by students who self-identify as confident academic writers versus those who do not. These resources provide useful data for beginning to address a number of questions that emerge from the issue of confidence in DSP, and in writing more generally.

One way to understand more about the nature and function of confidence is to consider its relationship to competence in writing. Our preliminary investigation of the relationship between student confidence and competence has focused on features that research shows to correlate with highly ranked writing. Two features emerge directly from the genre of writing required by the DSP prompt. One is organization, and we can learn something about this from analyzing the corpus for discourse markers such as transition words, since such markers correlate highly with effective argumentative writing (Xing et al., 2008). Another is reference to the reading material because research (Woodward-Kron, 2003) shows the importance of interacting with multiple voices to make effective arguments, and examples from the text are one of the features mentioned in the DSP prompt.

In addition, there are features that correlate with effective writing more generally. One of these is text length because research shows that students who produce more words typically receive higher scores, particularly on timed writing tests (Friedlander, 1990). Another is type/token ratios because research shows that students who use a greater variety of words are typically identified as better writers (Engber, 1993).

We believe that analyzing the entire corpus as well as subgroups identified by levels of self-proclaimed confidence for features like transition words and references to the reading material as well as text length and type-token ratios will provide some insight into the relationship between confidence and competence in writing. At the same time, however, we welcome discussion on how we might nuance this investigation further by calling upon other resources of computational linguistics.

4 Questions of Language Learning

As mentioned earlier, one of the subgroups within the larger university population is English Language Learners. Briefly, international students at our university who score below a fixed threshold on the TESOL are required to take a second test, the AEE, in addition to participating in the DSP process. The survey questions to which they respond are slightly different from those answered by native speakers, and the essay they read includes glosses to explain culturally specific terms. This combination of accommodations and measures is relatively effective in identifying students who need special intervention in order to write well in English.

But, as current research shows, there is another population of English Language Learners that is much less visible than the typical international students—the population typically known as Generation 1.5. These students are much more fully assimilated into US culture, usually because they have lived in this country for an extended period and have attended US schools. However, their writing frequently manifests many of the same difficulties as the more easily identified English Language Learners. One of the chief instructional challenges posed by Generation 1.5 students is that they are not easy to identify, and their instructional needs are not clearly defined. Analysis of the DSP process at our university shows that Generation 1.5 students regularly fly under the radar of self-placement and find themselves struggling in writing classes. Anecdotal reports from instructors point to these students' difficulties, but we have no systematic way of identifying and helping them. This population, like that of ELL

students, is currently growing each year, and it is becoming increasingly important to address its needs.

It is clear, however, that positioning English language learners and, especially, Generation 1.5 as deficient is not constructive. ELL and Generation 1.5 students are often constructed in highly positive terms such as hard-working and determined in high school and then positioned negatively as resistant and unmotivated when they enter college writing classes. The first challenge is to develop better ways of identifying Generation 1.5 students early in their university work so that they are not left to flounder, as they so often do, when they move into upper division courses. The double challenge of acquiring academic literacy while simultaneously acquiring proficiency in the English language frequently, as Short and Fitzsimmons (2007) show, becomes overwhelming to students who have many competencies and are highly motivated. The college writing class offers a space for equipping students who are learning English at the same time that they are leaning about college writing. In order for this to happen, however, we need to learn more about the specific nature of challenges faced by these students. Research by Wu (2007) shows that ability to adjust dialogic space is often difficult for L2 writers, and Hyland and Milton (1997) demonstrate that L2 writers frequently take a more authoritative and less nuanced stance, while more highly valued writing typically expresses more epistemic uncertainty.

As a first step, we will create a sub corpus of identified English Language Learners and use the rhetorical and interactive features of competence (organization, reference to reading, text length, lexical variety, and transition words) identified above to determine the extent to which these features identify levels of writing competence for this population. If we can isolate features that are characteristic of this population of English Language Learners, then we can attempt to apply the same features to the entire corpus in order to begin the process of identifying Generation 1.5 students.

We are less certain about how to use computational linguistics most effectively to identify ability to adjust dialogic space and take a more nuanced stance in writing. Nor, of course, are

we certain that these features will be the most productive in helping us to identify Generation 1.5 students. Accordingly, we will welcome discussion of additional ways to use computational linguistics to identify Generation 1.5 students.

5 Conclusion

We have done some preliminary thinking and begun investigations of questions about validity, confidence and English Language Learners, and we welcome the opportunity to explore ways of uniting research in writing and in computational linguistics to further our investigation.

References

- C. Engber. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4(2):139–155.
- A. Friedlander. 1990. Composing in English: Effects of a first language on writing in English as a second language In Barbara Kroll (Ed.) *Second Language Writing: Research Insights for the Classroom*, New York: Cambridge UP.
- A. R. Gere, L. Aull, T. Green and A. Porter. (forthcoming). Assessing the validity of directed self-placement at a large university. *Assessing Writing*.
- K. Hyland. 2005. Representing readers in writing: Student and expert practices. *Linguistics and Education* 16, 363–377.
- K. Hyland, and J. Milton. 1997. Qualifications and certainty in L1 and L2 students writing. *Journal of Second Language Writing* 6(2):183–205.
- E. J. Reynolds. 2003. The role of self-efficacy in writing and directed self-placement. In Daniel Royer and Roger Gilles (Eds.) *Directed Self-Placement: Principles and Practices*. Cresskill, NJ: Hampton Press, 73–104.
- R. Woodward-Kron. 2003. Critical analysis and the journal article review assignment *Journal of English for Academic Purposes*, 6, 254–271.
- M. Xing, J. Wang and K. Spencer. 2008. Raising Students' Awareness of cross-cultural contrastive rhetoric in English writing via an E-learning course. *Language Learning & Technology* 12(2):71–93.