Prototyping a Learning Analytics Tool to Explore Academic Language Development in Massive Open Online Courses

Challenge.
With massive open online courses (MOOCs) and other virtual learning environments, we are collecting vast amounts of fine-grained data on student activity and achievement. What are the meaningful signals of learning among this data? How can we represent and understand the complexity of learning by visualizing multiple dimensions of student performance?

Motivation.
Education is becoming increasingly digitized. Virtual environments from MOOCs to the Khan Academy to educational software in classrooms are recording each move that a learner makes. Learning analytics is developing as a field to harness this influx of “big data,” but because learning is a complex psychological and social phenomenon, there are many open questions about how to interpret this data to refine theories of how people learn. Learning analytics researchers need tools that will allow us to easily explore relationships between multiple dimensions of learner data.

We develop a prototype for this work by narrowing in on two salient indicators of learning from the Introduction to Mathematical Thinking MOOC on Coursera: test scores and the development of domain-specific, academic language. Research in the learning sciences shows that the development of domain-specific, academic language is a powerful indicator of learning, but this aspect of learning has been largely unexplored in the limited number of “dashboards” for virtual learning environments that have been developed to date.

Our visualization is designed to answer the particular research question: how does the adoption of academic language correspond to instructor language use and student performance on course assessments?

Approach.
We explored several possible designs before deciding on a stacked graph that captured multiple metrics of interest to our research in a single screen. Our implementation uses D3, with data transformation and text processing done primarily in Python. For simplicity of interpretation, we only show assessment scores for highly engaged students who completed all of the problem sets. The Pretest and Posttest scores are from a survey that asks students to self-report how relevant math is to them.

Future Work.
Systematizing and generalizing the tool
We will allow users to dynamically query terms of interest by integrating our Python data analysis code more fully with D3 in a Django web environment - something we did not have sufficient prior knowledge nor learning time to do for this prototype. We also plan to make the system adaptable to a range of performance data – i.e. different types of assessments and other indicators of engagement. We will also ensure compatibility with data from other courses within Coursera as well as other online education platforms.

Refining underlying data
We use a relatively basic indicator of language use – timing and frequency of particular terms. For a deeper understanding of learning we will need to examine language in context – for example, whether terms are being used formally or informally, accurately or inaccurately, inquisitively or declaratively. We plan to develop significantly stronger competency in Natural Language Processing in order to do this work, and will revise our visualization system to represent the most salient indicators of learning and engagement that emerge from these future analyses.

User testing
As learning analytics researchers we are our own pilot group, but we also plan to share the tool with other researchers to understand how it could help them answer their own research questions, and refine our features accordingly. We will also explore the usefulness of this tool for MOOC instructors.