Quantext: Analysing student responses to short-answer questions

Jenny McDonald
University of Auckland

Adon Christian Michael Moskal
Otago Polytechnic

We introduce a web-based tool for teachers to support the rapid analysis of student responses to short answer or mini-essay questions. Designed to support teaching in large-class settings, it aims to bring to practicing teachers analytic tools that can reveal insights in their student text data. We background development of the tool to date, briefly describe its architecture and features, and report on a bench-test evaluation. Finally, we introduce a pilot study to evaluate the tool in classrooms at three NZ universities and one polytechnic. We conclude with options for accessing the tool and outline plans for ongoing development.

Background

Quantext— a text analysis tool for teachers—has grown out of a demonstrable need, especially in large class settings, to rapidly evaluate written student responses to short-answer questions (McDonald, Bird, Zouaq & Moskal, 2017). Student success in higher education is predicated on interpreting, synthesising and producing text within specific disciplinary contexts. Yet, despite generating enormous volumes of text with each student cohort, the current preoccupation of learning analytics is with proxies of student engagement and learning; for example, counting clicks to access course materials, counting assignments completed, or ranking test scores (Ferguson, Brasher, Clow, et al., 2016). This project takes a different approach; we focus firmly on the words of the students themselves; arguably, the ‘site of learning’ (Knight & Littleton, 2015 p.).

Even though analysis and synthesis of text are central to teaching and learning in higher education, and even though this work is conceptually difficult, little attention is paid to how students understand and interpret teacher language in constructing their own academic writing (Laurillard, 1993). While there is certainly educational research around how students come to understand academic discourse (e.g. Marton & Säljö, 1976), around teaching academic writing (Lea & Street, 1998), and around the link between language and learning (e.g. Gee, 2015; Wells, 1994; Halliday, 1993), translating this research into actionable insights for teachers, particularly of larger classes, remains elusive.

However, this is not for want of data. Both student writing and text-based teaching materials are routinely uploaded to institutional Learning Management Systems (LMS). These data are used for assessment purposes or checking for plagiarism but are rarely consulted in systematic ways for improving teaching or informing learning design. We argue that it is essential that we not overlook the opportunity to analyse these data to illuminate student learning.

Furthermore, from a dialogic perspective (Bakhtin, 1981), what we write or speak about, is intimately related to what we have read or listened to. Therefore, our analysis must also include the teacher and teaching materials or we will fail to capture the dialogic at the centre of teaching and learning. In short, we suggest that analysis of student text must go together with the analysis of teacher text.

Our goal in developing Quantext is to bring to practicing teachers analytic tools that utilise the vast quantities of text data already being collected, as well as facilitate the analysis of text in settings, such as large classes, where this is currently impractical. Analysis of these data should expose and illuminate the site of learning, ultimately enhancing both teaching and learning.

Quantext development and prototyping

Our concept developed from a case study exploring student text responses to short-answer questions in the context of a large first year health sciences course (McDonald, Bird, Zouaq & Moskal, 2017). The case study was part of a larger NZ-wide learning analytics project (Gunn et al., 2016) funded by Ako Aotearoa (NPF15-008). This case study revealed multiple relationships between student responses, course materials and questions asked. We concluded that a tool, based on established methods of corpus linguistics and natural language processing (hereafter referred to broadly as text analysis), could provide timely, actionable insights for teachers and help foster deep learning approaches for students.

As part of Ako NPF15-008, we held a total of four workshops at NZ-tertiary institutions during 2016 to introduce text analysis tools and approaches to teachers. While there was interest and enthusiasm from workshop participants, most existing text analysis tools were beyond the reach of most practising teachers. Furthermore, even getting text data into a form suitable for analysis presented challenge.

The challenges identified in the workshops helped define what the key requirements for Quantext should be: i) the tool should be available and accessible online; ii) uploading text data should be straightforward and eventually integrated with student data held in institutional LMS; iii) no prior knowledge of linguistic terms or metrics should be assumed; iv) interface tools should use familiar analysis paradigms (e.g. spreadsheets) and basic charts/visualisations; v) workflow should be straightforward and result in a specific output (e.g. label responses with teacher-defined categories or marking rubric); vi) text analysis settings should be accessible and easily customisable as skill with the tool develops; and vii) the tool should enable insights which can form the basis of feedback to students, and inform learning design and teacher development.

An initial prototype was developed following informal discussions with specialist academic developers at the University of Auckland and Victoria University of Wellington, as well as from interested tertiary teachers at several NZ institutions. Quantext is currently at the minimum viable product (MVP) stage; that is, some aspects are incomplete, but there is sufficient functionality for teachers to assess its suitability for classroom use (Münch et al., 2013). We describe the key features and workflow below.

Key features and workflow

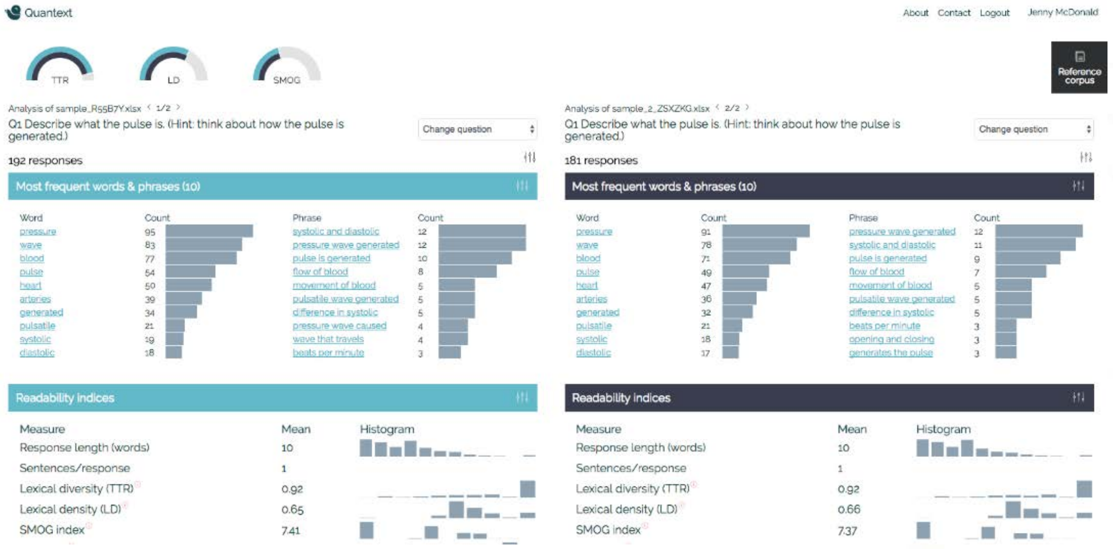
There are multiple, often conflicting approaches to evaluating student responses to short answer questions (whether formative or summative). As Orr (2007) points out, the territory between positivist and poststructural approaches to assessment is complex and multilayered. Nevertheless, there is little recognition of this complexity in the computational assessment literature, which almost always evaluates automated methods of assessment against ‘gold-standard’ human markers. Typically, if interrater reliability between an automated marker and human marker is comparable to the interrater score between human markers, then the automated marker is performing well. Note, however, that interrater scores between humans, can be highly variable, due in part to the complexity of the assessment landscape (Jonsson & Svingby, 2007). To be fair to those working in computational assessment, while teachers may operate anywhere along the epistemological spectrum, in practice, institutional, disciplinary and curricula constraints combine to result in practical approaches to assessment characterised by a distinctly positivist stance. It is hardly surprising then that automated evaluation of short-answer questions emphasises standards and measurement.

In contemporary undergraduate classes, if evaluation of short-answer question responses occurs at all (in large classes, students may simply be given model answers to self-assess), it typically involves one of these approaches: i) a binary approach to evaluate whether responses are the same or different to a model/reference response; ii) a grading approach where a marking rubric is applied to evaluate whether all or some components of a model answer are present; or iii) a best judgement approach where the rater simply allocates grades or marks (although interrater checks may be made to ensure consistency across a cohort). With each approach, the rater may be a teacher, a tutor, a peer, or a machine.

Because of the complexity of the assessment space, response evaluation in Quantext deliberately makes no assumptions about the specific evaluation approach or epistemological stance. For example, the similarity metric may be used with a model answer, a representative mis(conception), or another student response—the choice of reference response is up to the teacher. Quantext simply compares the reference to each student response, and returns a number between 1 to -1: student responses sharing linguistic and semantic features with the reference response score closer to 1; responses unalike score closer to 0; and responses completely opposite score closer to -1. The teacher can sort all responses by this similarity metric to find those most similar, and label/categorise them accordingly.

The Quantext workflow is: i) upload questions and responses in spreadsheet format; ii) select which responses to analyse (you can choose more than one dataset for comparing different student cohorts); and iii) run the analysis. Default analysis provides descriptive statistics and charts for each dataset, including number of responses, length of response (word or sentence count), and readability indices (e.g. lexical diversity and lexical density). There is also a customisable keyword/key phrase display. By default, keywords are a frequency count of the most commonly occurring words excluding stopwords (i.e. functional words like ‘the’, ‘and’, ‘of’, etc.), and key phrases are word pairs (bigrams) or triples (trigrams) which occur together more commonly than by chance. Finally, there is a worksheet view of all student responses along with derived descriptive statistics (number of words, lexical diversity, etc.), and similarity to reference response (if given). The worksheet is searchable, easily filtered and sortable on any column.

Responses can be filtered to show only those containing a selected word or phrase. A label tool allows teachers to define categories for any student response. Figure 1 shows a screenshot of the Quantext analysis screen.



Analysis of sample_R55B7Y.xlsx ‹ 1/2 ›

Q1 Describe what the pulse is. (Hint: think about how the pulse is generated.)

192 responses

Most frequent words & phrases (10)

Word	Count	Phrase	Count
pressure	95	systolic and diastolic	12
wave	83	pressure wave generated	12
blood	77	pulse is generated	10
pulse	54	flow of blood	8
heart	50	movement of blood	5
arteries	39	pulsatile wave generated	5
generated	34	difference in systolic	5
pulsatile	21	pressure wave caused	4
systolic	19	wave that travels	4
diastolic	18	beats per minute	3

Readability indices

Measure	Mean	Histogram
Response length (words)	10	
Sentences/response	1	
Lexical diversity (TTR)	0.92	
Lexical density (LD)	0.65	
SMOG index	7.41	

Analysis of sample_2_ZSXZKG.xlsx ‹ 2/2 ›

Q1 Describe what the pulse is. (Hint: think about how the pulse is generated.)

181 responses

Most frequent words & phrases (10)

Word	Count	Phrase	Count
pressure	91	pressure wave generated	12
wave	78	systolic and diastolic	11
blood	71	pulse is generated	9
pulse	49	flow of blood	7
heart	47	movement of blood	5
arteries	36	pulsatile wave generated	5
generated	32	difference in systolic	5
pulsatile	21	beats per minute	3
systolic	18	opening and closing	3
diastolic	17	generates the pulse	3

Readability indices

Measure	Mean	Histogram
Response length (words)	10	
Sentences/response	1	
Lexical diversity (TTR)	0.92	
Lexical density (LD)	0.66	
SMOG index	7.37	

Figure 1: Quantext analysis: most frequent words, phrases and readability indices of two datasets

Through filtering and sorting the responses by the basic indices, word, bigram and trigram frequency, and similarity, we anticipate that teachers will be able to rapidly evaluate and categorise large numbers of student text responses. The resulting analysis can also be exported for comparison with other student data.

Teachers also have the option to augment their analysis through uploading teaching material related to the questions being asked, which serves two key functions: i) this provides a reference corpus with respect to student responses (e.g., one might anticipate key words and pairs from the teaching corpus to appear in the student responses and vice versa), facilitating identification of pedagogenic (mis)conceptions (Laurillard, 2002) through a keyword-in-context display; and ii) the readability indices of the teaching corpus can be checked against the readability indices of the student responses.

Bench-test baseline evaluation

To assess potential benefits to teachers and students, evaluation of Quantext in authentic educational settings is planned (we describe a forthcoming pilot study below). However, a key feature of the Quantext workflow, as mentioned above, is measuring the similarity of student responses to a reference response. We conducted a baseline evaluation of this feature ahead of the pilot study, as a 'bench-test', using a dataset of 924 student responses to 10 open questions (in McDonald, Bird, Zouaq & Moskal, 2017). The questions related to a first-year undergraduate health sciences programme and were relational or multi-structural in nature (see SOLO Taxonomy, Biggs & Collis, 1982). In other words, they went beyond testing simple recall of facts to ask deeper questions. All student responses were labelled by two human markers who negotiated the appropriate label/s. It is important to note that more than one label could apply to any given response. In assessing similarity[1], for the purposes of our bench test, a single human assigned label was chosen for each response and compared to a reference response with the same label. A summary of our results for the 'correct' label is presented in Table 1.

[1] Similarity is calculated from a word2vec model of word embeddings using the GloVe algorithm (Pennington, Socher & Manning, 2014) and is pre-trained on the Common Crawl Corpus (Spiegler, 2013). An average response vector is calculated from the word vectors in each response and then the cosine distance between the response vector and the reference response is computed to give a similarity score between 1 and -1. Quantext uses the Spacy library (https://spacy.io) for the pre-trained word2vec model.

Table 1: Labelled responses with similarity measure of >= 0.90 to reference answer, ’correct

Question	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
True Positive	15	19	3	2	3	2	12	14	8	14
False Negative	2	25	5	9	3	22	2	1	12	5
True Negative	138	95	108	93	69	50	36	36	27	11
False Positive	37	4	11	1	1	0	15	11	0	3
Accuracy	0.8	0.8	0.87	0.9	0.95	0.7	0.74	0.81	0.74	0.75
Recall	0.88	0.43	0.37	0.18	0.5	0.08	0.86	0.93	0.4	0.74
Precision	0.29	0.83	0.21	0.66	0.75	1	0.44	0.56	1	0.82

Evaluation with other labels (e.g. ‘incomplete’, ‘don’t-know’, ‘incorrect’, ‘naïve’ etc) produced similar results. While far from perfect, we believe that overall accuracy of 0.70–0.95 for an out-of-the-box similarity algorithm on open-ended questions is acceptable for computer assisted evaluation. In particular, we suggest this is the case when other options to support label assignment such as keywords and response length are available to the teacher. One goal of the pilot study will be to test this belief through adopting a similar approach to Basu, Jacobs and Vanderwende (2013). We also hope to improve on baseline similarity performance through augmenting model training on domain-specific corpora.

Pilot study

We have recruited interested teachers from four NZ tertiary institutions to pilot Quantext in semester two (starting July, 2017)—at the time of writing, five teachers from five different courses, spanning the humanities, sciences, health sciences and commerce, and with cohorts of under 100 students to over 1000. All courses are on-site rather than distance courses. This reflects the courses taught by pilot volunteers rather than a deliberate choice. Specific considerations to explore in the pilot study are:

1. Evaluate the utility of Quantext, along the following dimensions to support student learning and engagement: (i) utility and accessibility of the tool; ii) validity and reliability of data; iii) intended vs actual use of the tool; iv) identification of actionable teaching insights; v) identification of insights to improve course design; and vi) utility at different stages of the teaching/learning design cycle).
2. Though teacher reflections, explore the impact, if any, of tool use on: i) student learning; ii) participant teaching practice; and iii) participant professional development.

Broadly, the pilot will adopt a development or design-based research approach. This means we will treat each course in the pilot as an individual case study. The pilot will begin with an introductory seminar/workshop at each site, covering administration details, and planning the questions to be asked of students. This will include discussing frameworks about the framing and motivation for asking short-answer questions (e.g. SOLO taxonomy), and addressing ethical or operational issues.

For the duration of the pilot, participants will ask formative, open-ended questions of their students and use Quantext to help evaluate student responses. It will be entirely up to participating teachers how they choose to incorporate short-answer questions for analysis with Quantext into their course. Examples include: i) questions may be asked at the start and again at the end of the course/module to see if there are changes or development in student language; or ii) questions may be asked at any time throughout the teaching period to explore emerging student understanding of specific concepts.

We envisage teachers will use existing systems such as Canvas, Blackboard, Moodle or similar, to facilitate collecting student responses in digital form. Relevant teaching materials will also be uploaded as reference corpora for the student responses. Ideally a complete set of teaching materials will be used, such as lecture notes, transcripts or textbooks (although some material may be not be available for inclusion), and we will assist teachers with creating their reference corpora.

Teachers using Quantext to analyse student responses will evaluate their analyses according to pilot goals. Throughout the pilot, teachers are welcome to give feedback or seek advice from the pilot project team, and the developers will be available to fix and update the software as problems are identified. We will capture teacher analyses conducted using the tool to form part of the pilot dataset. Concluding focus group sessions with teachers will be held at each site at the end of semester two.

Conclusion

We introduce a novel, web-based tool for teachers to support the rapid analysis of student responses to short answer or mini-essay questions. Results from a planned NZ pilot study will inform ongoing tool development. We hope to present early results from the pilot during Ascilite 2017. An evaluation version of Quantext and a link to the source code hosted on Github is available at http://www.quantext.org

References

Bakhtin, M. M. (1981). *The dialogic imagination*. Austin: University of Texas Press.

Basu, S., Jacobs, C & Vanderwende, L. (2013) Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, 1 (2013) 391–402.

Ferguson, R., Brasher, A., Clow, D., Cooper, A., Hillaire, G., Mittelmeier, J., Rienties, B., Ullmann, T. and Vuorikari, R. (2016). *Research Evidence on the Use of Learning Analytics: Implications for Education Policy*. Joint Research Centre, Seville, Spain

Gee, J. (2015). Social linguistics and literacies: Ideology in discourses: Routledge.

Gunn, C., Blumenstein, M., Donald, C., McDonald, J., & Milne, J. (2016). *The Missing Link for Learning from Analytics.* In Show me the learning: Ascilite 2016. Barker, S. Dawson, A. Pardo, & C. Colvin (Eds), Adelaide.

Halliday, M. A. K. (1993). Towards a language-based theory of learning. *Linguistics and Education*, 5(2), 93–116 .http://doi. org/10.1016/0898-5898(93)90026-7

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational research review*, *2*(2), 130-144.

Knight, S., & Littleton, K. (2015). Discourse-centric learning analytics: mapping the terrain. *Journal of Learning Analytics*, *2*(1), 185-209.

Laurillard, D. (1993) Rethinking University Teaching: a framework for the effective use of educational technology. Routledge, London.

Laurillard, D. (2002) Rethinking University Teaching: a framework for the effective use of educational technology. RoutledgeFalmer, London.

Lea, M. & Street, B (1998) Student writing in higher education: An academic literacies approach. *Studies in Higher Education.* 23(2), 157-172

Marton, F. and Saljo, R. (1976). On qualitative differences in learning: I Outcome and process. British Journal of Educational Psychology, 46(1), 4–11.

McDonald, J., Bird, R. J., Zouaq, A., & Moskal, A. C. M. (2017). Short answers to deep questions: supporting teachers in large-class settings. *Journal of Computer Assisted Learning*. 33(4) 306–319. doi: 10.1111/jcal.12178.

Münch J., Fagerholm F., Johnson P., Pirttilahti J., Torkkel J., Järvinen J. (2013) Creating Minimum Viable Products in Industry-Academia Collaborations. In: Fitzgerald B., Conboy K., Power K., Valerdi R., Morgan L., Stol KJ. (eds) Lean Enterprise Software and Systems. Lecture Notes in Business Information Processing, vol 167. Springer, Berlin, Heidelberg

Orr, S (2007) Assessment moderation: constructing the marks and constructing the students, Assessment & Evaluation in Higher Education, 32(6), 645-656, DOI: 10.1080/02602930601117068

Pennington, J., Socher, R., & Manning, C.D. (2014) GloVe: Global Vectors for Word Representation. In proceedings, *Empirical Methods in Natural Language Processing (EMNLP)*. pp. 1532-1543

Spiegler, S (2013) Statistics of the Common Crawl Corpus. Retreived from http://bit.ly/2rRWsX9 4th June, 2017.

Wells, G. (1994). The complementary contributions of Halliday and Vygotsky to a “language-based theory of learning”. Linguistics and Education, 6(1), pp. 41–90.

Contact author: Jenny McDonald, j.mcdonald@auckland.ac.nz .
Please cite as: McDonald, J. & Moskal A.C.M. (2017). Quantext: analysing student responses to short-answer questions. In H. Partridge, K. Davis, & J. Thomas. (Eds.), *Me, Us, IT! Proceedings ASCILITE2017: 34th International Conference on Innovation, Practice and Research in the Use of Educational Technologies in Tertiary Education* (pp. 133-137).

Note: All published papers are refereed, having undergone a double-blind peer-review process.