Using student experience as a model for designing an automatic feedback system for short essays

An exploration of the features of graded student essays using domain-independent natural language processing techniques

Debora Field[1], John T. E. Richardson[2], Stephen Pulman[1], Nicolas Van Labeke[2] and Denise Whitelock[2], [1]University of Oxford and [2]The Open University

Abstract
This paper presents observations that were made about a corpus of 135 graded student essays by analysing them with a computer program that we are designing to provide automated formative feedback on draft essays. In order to provide individualised feedback to help students to improve their essays, the program carries out automatic essay structure recognition and uses domain-independent graph-based ranking techniques to derive extractive summaries. These procedures generate data concerning an essay's organisational structure and its discourse structure. We have selected 27 attributes from the data and used them in a comparative analysis of all the essays with a view to informing further development of the feedback program. The results of this analysis suggest that some characteristics of students' essays that our domain-independent feedback program is measuring may be related to the grades that tutors assign to their essays.

Keywords
Graph-based ranking; key sentence extraction; key word extraction; natural language processing; students' essays.

Introduction

In recent years there has been increasing recognition that assessment is not simply a means of confirming that learning has taken place (in other words, assessment *of* learning) but that it can also help to reinforce the process of learning itself (in other words, assessment *for* learning). For the latter to be achieved, assessment has to be accompanied by appropriate and meaningful feedback. A number of authors have discussed the conditions under which feedback is effective in promoting learning (Evans 2013; Gibbs and Simpson 2004–05; Hattie and Timperley 2007; Nicol and Macfarlane-Dick 2006).

Following Chickering and Gamson (1987, 1999), several of these authors stress that feedback needs to be timely. It needs to be 'received by students while it still matters to them and in time for them to pay attention to further learning or receive further assistance' (Gibbs and Simpson 2004–05, 18). Nowadays, this may be extremely hard for academic staff to achieve with increasing class sizes and additional factors that serve to increase their workload. Perhaps electronic assessment systems can provide timely and meaningful feedback when their teachers are unable to do so. Conversely, electronic forms of assessment in higher education need to incorporate appropriate kinds of feedback so that students can understand why their work is being assessed in the way that it is and consequently can learn from the experience (DiBattista, Mitterer, and Gosse 2004; Pitcher, Goldfinch, and Beevers 2002; Walker, Topping, and Rodrigues 2008; Whitelock and Raw 2003). This will allow students to take more control of their learning, become more reflective and improve their learning skills (Whitelock and Brasher 2006).

The SAFeSEA project (Supported Automated Feedback for Short Essay Answers) aims to develop an automated system to provide students with helpful and constructive feedback on their draft essays. Such support requires research into the selection of the content and the mode of presentation and delivery of the feedback. For the feedback to be effective, students need to be helped both to manage their current essay-writing task and to develop their essay-writing skills. The main thrust of the project involves using state-of-the-art techniques in natural language processing to analyse students' essays, developing a range of feedback models and evaluating their effectiveness.

Educational research suggests that one particular type of feedback that falls within the scope of natural language processing – essay summarisation – is among the most useful for students (Nelson and Schunn 2009). 'Summarisation' includes both the traditional notion of a short précis and also simpler representations such as a list of an essay's key topics. As part of a larger prototype application, we have implemented essay structure recognition and key word and key sentence extraction procedures in a module that we call 'EssayAnalyser'. We have used the module to explore the attributes of a corpus of 135 essays (with a word limit of 1,500 words) that were produced by students taking a postgraduate course by distance learning. This paper describes the module's design and the results of our exploration.

Graph-based ranking methods

Our procedures are based on graph theory, which has been used in a wide variety of disciplinary contexts. The following account is based on that provided by Newman (2008). A graph consists of a set of *nodes* or *vertices* and a set of *links* or *edges* connecting them. (Some writers describe such a system as a *network*, but others restrict the latter to refer to graphs in which the edges are both directed and labelled.) A graph can be represented by a matrix of adjacencies in which the cells represent the connections between all pairs of nodes. In the simplest case, the cells take the value 1 if there is an edge between the relevant nodes and 0 otherwise.

Measures of *centrality* identify the most important or central nodes in a graph. They can therefore be used to measure how central (or key) a word, phrase, or sentence is in a natural language text of arbitrary length. The simplest such measure is *degree*, which is simply the number of edges attached to a node. Some other centrality measures take into account how strongly connected each node in the graph is to the whole graph, rather than just to its neighbouring nodes. We have used two of the latter centrality measures: eigenvector centrality (Brin and Page 1998) and betweenness centrality (Freeman 1977). (It should be noted that our approach is very different from latent semantic analysis).

Text pre-processing and essay structure recognition

Before extracting key words and sentences from the text, the text is automatically pre-processed using modules from the Natural Language Processing Toolkit (Bird, Klein, and Loper 2009). We remove 'stop words' (articles, prepositions, auxiliary verbs, pronouns, etc.), which are the most frequently occurring in natural language but for our purposes the least interesting. We refer to the remaining meaning-rich words as 'tidy' words and to the sentences without stop words as 'tidied' sentences.

Structural components present in the essay are also automatically recognised and labelled (currently including preface, summary, abstract, introduction, discussion, conclusion, table of contents, title, captions, list items, table entries, headings, quoted assignment question sentences, references, and appendices). This enables us to choose the sections of the essay that we wish to analyse for the presence of key sentences and key words, and it also allows us to highlight different parts of the essay in the feedback application's visual representations. We consider the prose of the body of the essay (introduction, discussion and conclusion) to be the most suitable source material for both key words and key sentences, and so the other essay parts are omitted from the summarisation procedures. We refer to the essay's prose sentences as 'true sentences'.

As instructions in the assessment task concerning essay structure were minimal (see later), and as the resources and applications used by the students to compose the essays had varied widely (since distance-learning students work remotely rather than on campus), the essays in the corpus vary greatly in structure and formatting choices that impact on structure. It was therefore decided that structure recognition would be best achieved without referring to a high-level formatting mark-up, and so the essays are converted to plain text files in UTF-8 encoding before they are processed by EssayAnalyser. The structure recognition rules have been hand-crafted from extensive experimentation with the corpus.

Key word extraction

After text pre-processing, graph-based ranking methods are used to ascribe a 'key-ness' rank to the lemma (or base form) of each word in an essay. This follows Mihalcea and Tarau (2004), except that we use betweenness centrality to measure the centrality of a lemma in a text rather than eigenvector centrality. Key lemmas are defined as those in the top 20% of the ranked nodes that have betweenness centrality scores of .03 or more. (This threshold is where visual inspection identifies the sharpest bend in the 'elbow' of the distribution curve in the key word centrality scores across all of the essays.) The essay's key words are the inflections or base forms of the key lemmas that occur in the essay's original text. Key phrases are within-sentence sequences of key words that occur in the original text.

A visualisation of the key word graph for a very short sample text is shown in Figure 1. Each lemma in the text (also shown) is represented by a node in the graph (a large dot) and lemmas whose surface forms are adjacent in the text are connected by edges (lines). In the key word graph, a node's centrality is defined as "the degree to which a point falls on the shortest path between others" (Freeman 1977, 35).

Key sentence extraction

Key sentences are also extracted using a graph-based ranking method. Instead of lemmas, every *sentence* in the essay is represented by a node in the graph. Each true sentence is compared with every other true sentence, and a value is derived representing the semantic similarity of that pair. That similarity value becomes a weight that attaches to the edge that links the corresponding nodes in the key sentence graph. We are currently using cosine similarity as the similarity measure. The nodes are ranked using Mihalcea and Tarau's (2004) TextRank algorithm, and key sentences are defined as the top 30 ranked sentences. Note that no domain knowledge, other expert knowledge or 'gold standard' model specific to a particular

domain is used in the program's extraction of key words and key sentences.
An annotated illustration of the key sentence graph for the same very short text is shown in Figure 2.

"Will the foxes in my garden attack my dog or cat?

"This is extremely unlikely. Foxes avoid dogs, even small dogs, because many foxes are killed by dogs. So it is much more likely that your dog will attack the fox, not the other way round. Attacks on cats are equally rare: cats and foxes are roughly the same size, and cats are very capable of defending themselves against foxes. So it is hardly surprising that foxes generally give cats a wide berth and flee when threatened by a cat. Occasionally small kittens are killed, but this is rare. Keeping your cat indoors at night greatly reduces the chances of an encounter with a fox. There are also a variety of other benefits: cats kept in at night are healthier and live longer, and kill less of the local wildlife."

Source text from:
http://www.thefoxwebsite.org/faq/urbanfoxdisease.html#q4

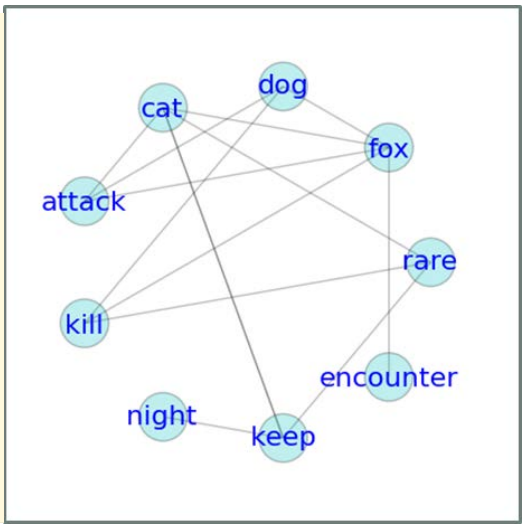

Figure 1: Illustrative small text with key words underlined and visualised key lemma graph (right)

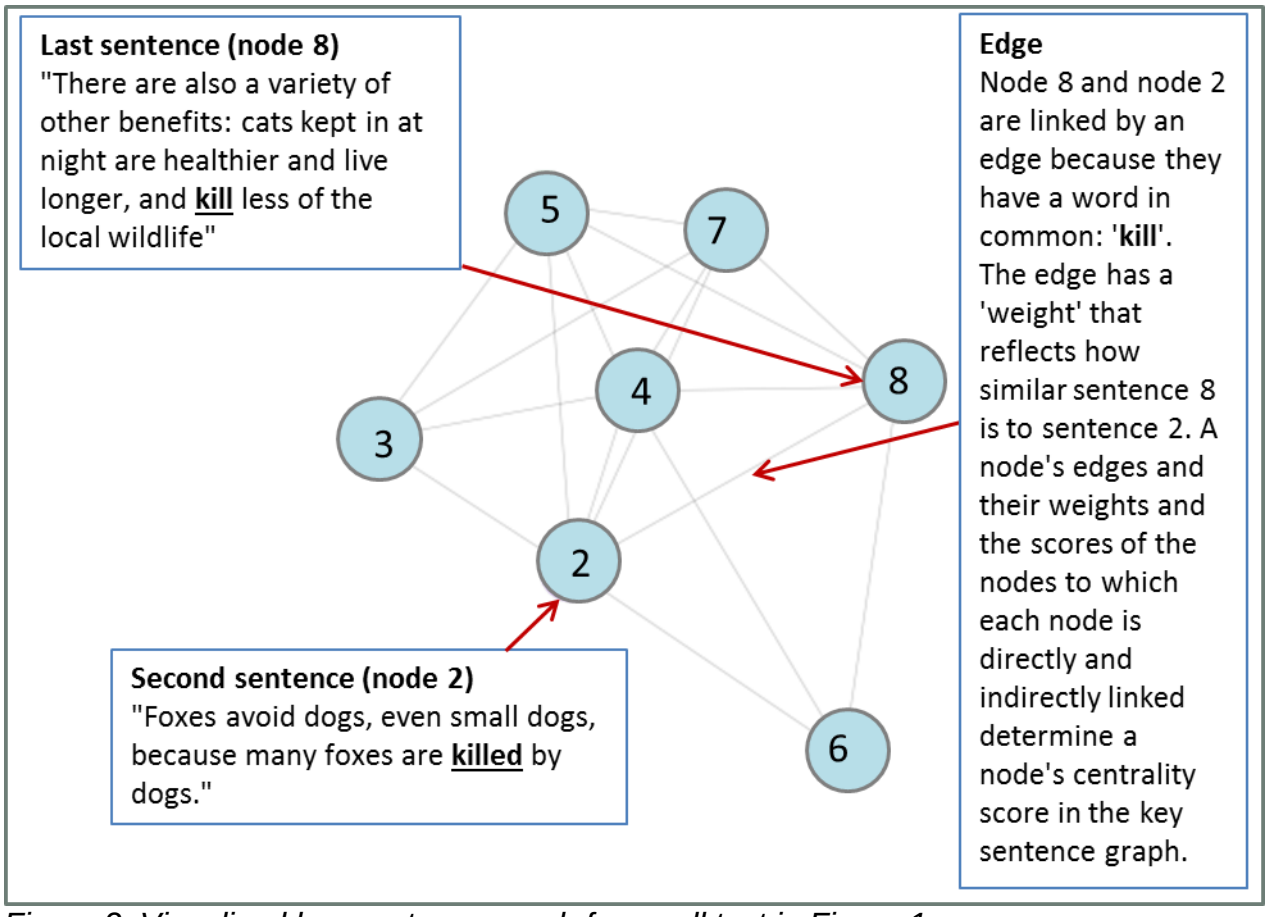

Figure 2: Visualised key sentence graph for small text in Figure 1

Context
The essays were written by students taking The Open University's course H810, entitled *Accessible online learning: Supporting disabled students*. This postgraduate course is presented annually over 20 weeks between September and January. It is worth 30 credit points (and thus equates to one quarter of a year's full-time study). The course is supported by a textbook (Seale 2006) and by online resources, including links to a large number of external websites. Students are assigned to online tutorial groups and communicate with their tutors and one another through online forums. They are assessed by two assignments that are marked by their tutors and an end-of-course assignment that is marked by their tutor and an independent marker (another tutor). All assignments are marked using a percentage scale on which the pass mark is 40%.

Assessment task
The first assignment is submitted online six weeks after the beginning of the course. The task requires that students discuss accessibility challenges for disabled learners in the students' own work context. Many students are professionals with extensive work experience in a wide variety of areas. This means that, although there is a set course textbook, student essays vary greatly in subject matter.

Assessment criteria
The assessment criteria were stated as follows:

1. The extent and quality of your contributions to the tutor group forums during Block even a simple question, for example, or a straightforward suggestion, can make an extremely effective contribution to the discussion. This could include references to discussions in the general forum. (10 per cent)

2. The quality of your practical insights into the distinctive features of your chosen educational context. (25 per cent)

3. Your choice of online resources and the way in which you use them to support your argument. Your own argument still needs to be visible to the reader; some students weigh down nearly every sentence with a reference, which makes it hard to see what they are trying to say. (30 per cent)

4. The clarity, coherence and critical reflection of your written argument, showing that you have considered different points of view and then developed and clearly stated your own argument. (25 per cent)

5. Your written presentation: accurate use of the author/date referencing system, coherent style, choice of vocabulary, good grammar and accurate spelling. (10 per cent)

A total of 135 students submitted the first assignment in 2010, 2011 and 2012. The EssayAnalyser program generated 27 characteristics of these 135 essays. These 27 attributes are listed in Table 1.

Exploratory factor analysis

The purpose of factor analysis is to look for patterns in the relationships among a set of variables. If several variables are all very highly correlated with one another, then it is reasonable to assume that they are all tapping the same underlying construct. For example, if one asks a large sample of people which hand they use or prefer to use for different activities (e.g. writing, throwing, cutting with scissors, playing with a racket or bat, brushing their teeth and striking a match), the correlation coefficients among their responses will be positive and high, indicating that most people – but not all – report using the same hand for most of these activities. This suggests that it is sensible to talk about a single underlying dimension or factor of 'handedness' (Richardson 1978).

In some cases, it may be necessary or desirable to extract two or more underlying traits or factors. To aid in the interpretation of these factors, it is usually appropriate to transform or 'rotate' these factors within their *n*-dimensional space. The rotations in question may be *orthogonal* (which requires that the rotated factors should be perpendicular to one another) or *oblique* (which allows for the possibility that the rotated factors are correlated with one another).

An exploratory factor analysis was carried out on the values of these 27 attributes for the 135 essays. A sample size of 135 is lower than the minima recommended by traditional texts (e.g. Comrey 1973). However, more recent simulations have shown that robust results can be obtained from factor analyses with samples of 50–100 or fewer, even with large numbers of variables (de Winter, Dodou, and Wieringa 2009; Sapnas and Zeller 2002).

First, a principal components analysis was used to determine the number of factors to extract. This identified nine components with eigenvalues greater than 1, and these explained 83.5% of the variance in the data. Nevertheless, the eigenvalues-greater-than-one rule is known to overestimate the true number of components in a data set because of sampling effects (Cliff 1988). The bias is worse when the number of variables is large and the number of cases is small (both of which apply in the present case). Nowadays, it is generally acknowledged that the most reliable way to identify the number of factors in a data set is the parallel analysis of random correlation matrices. The analysis of 1,000 random correlation matrices was carried out using the program written by O'Connor (2000). The first seven components in the actual data set had eigenvalues greater than would be expected from a random data set, but the eighth and subsequent components did not. These seven components explained 74.7% of the variance in the data.

Principal axis factoring was therefore used to extract seven factors with squared multiple correlations as the initial estimates of communality, and the extracted factor matrix was submitted to oblique rotation using a quartimin method. A cut-off of $\pm$.50 was used to identify those loadings that were salient for the purposes of interpretation. In Table 2, the variables with salient loadings on each factor are listed in descending order of the loadings in question.

Table 1: Definitions of 27 attributes of students' essays (in alphabetical order)

Attribute name	Definition
% body == c	Percentage of the essay body (true sentences only) devoted to the conclusion section
% body == i	Percentage of the essay body (true sentences) devoted to the introduction section
all bigrams	Number of bigrams (made from key words)
all lemmas	Number of lemmas
all words	Number of words in the essay (occurring before the reference list or bibliography)
avfreq top5freq	Mean average frequency of the top five most frequent lemmas
avlen tidysent	Mean average length of a tidied sentence (a sentence without stop words in it)
bigrams in ass_q	Number of the essay's distinct bigrams that occur in the entire assignment question
c & toprank	Number of the top 30 key sentences that are in the conclusion section
distinct bigrams	Number of distinct bigrams
edges	Number of edges in the key sentence graph
edges/sents	Number of sentence graph edges divided by the number of true sentences
heads	Number of headings
i & toprank	Number of the top 30 key sentences that are in the introduction section
key lemmas	Number of key lemmas
key words	Number of key words
kls in ass_q_long	Number of essay's key lemmas occurring in whole assignment question
kls in ass_q_short	Number of essay's key lemmas occurring in assignment question's first sentence
kls in tb index	Number of essay's key lemmas occurring in course textbook index
len refs	Number of references in the references section
paras	Number of paragraphs
q sents	Number of sentences in whole assignment question quoted in the essay
sum freq kl_in_ass_q_long	Sum of the frequency counts (in the essay) for the essay's key lemmas that also occur in whole assignment question
sum freq kl_in_ass_q_short	Sum of the frequency counts for the essay's key lemmas that also occur in first sentence of assignment question
sum freq kls_in_tb_index	Sum of the frequency counts for the essay's key lemmas that also occur in the course textbook index
tidy words	Number of words in the essay ('all words') minus the stop words
true sents	Number of true sentences (excludes headings, captions, table of contents, title, etc.)

Table 2: Loadings of 27 variables on seven factors (with salient loadings in bold)

Attribute	1	2	3	4	5	6	7
avfreq top5freq	**.89**	.06	.04	-.02	.01	-.03	.09
edges/sents	**.89**	.05	-.18	.04	.04	-.04	.18
key lemmas	**-.88**	.04	.03	.01	.00	.01	.08
key words	**-.85**	.09	.03	.10	.07	.01	.05
edges	**.74**	.04	.27	.36	.01	-.04	.11
kls in tb index	**-.64**	-.01	.10	-.02	-.08	-.01	**.61**
sum freq kls_in_ass_q_short	.24	**.95**	.06	.00	-.01	.10	-.14
bigrams in ass_q	.07	**.78**	-.01	-.08	-.01	.01	-.05
kls in ass_q_short	-.12	**.72**	-.04	.06	-.04	-.03	-.08
all bigrams	.02	**.59**	-.02	-.14	.01	-.02	.27
sum freq kls_in_ass_q_long	.46	**.50**	-.03	.05	.00	.00	.41
kls in ass_q_long	-.25	.44	-.06	.03	-.06	-.02	.36
distinct bigrams	-.25	.40	-.09	-.01	.07	-.03	.27
q sents	-.09	.30	.06	.14	-.01	-.06	-.07
paras	-.08	-.02	**.89**	-.01	.03	-.02	.17
heads	-.07	.02	**.72**	-.10	.05	.01	.13
true sents	.11	.01	**.70**	.40	-.05	-.01	-.07
avlen tidysent	-.02	.02	**-.56**	.08	.01	.06	.19
all lemmas	-.29	-.06	-.11	**.86**	-.04	.00	-.12
all words	.27	-.06	-.11	**.84**	.05	-.01	.08
tidy words	.19	.08	.21	**.80**	.00	.03	.19
len refs	-.10	.12	.06	.28	-.14	.02	-.01
% body == i	-.04	-.01	.02	.05	**.98**	.02	-.02
i & toprank	-.06	.06	.05	.03	**.96**	.00	-.03
c & toprank	-.05	.03	.01	.03	.04	**.90**	.01
% body == c	-.01	.00	-.01	.00	-.03	**.89**	-.01
sum freq kls_in_tb_index	.22	-.07	.07	.07	-.02	.00	**.88**

The resulting solution exemplified ‘simple structure’ in that most of the variables loaded on one factor and only one variable loaded on more than one factor. The use of oblique rotation allowed for the possibility that the factors were correlated with one another. The correlation coefficient between Factor 2 and Factor 7 was .29. Otherwise, the correlation coefficients among the factors were all less than .20 in magnitude, implying that they were relatively orthogonal. It was therefore sensible to consider the variance explained by each factor.

Factor 1 explained 17.8% of the variance in the data set. Essays scored highly on this factor if (a) the frequency counts of the essay’s top five most frequent lemmas were high compared to other essays; (b) the number of edges in the sentence graph relative to the number of true sentences was high; (c) there were relatively few key lemmas; (d) there were relatively few key words; (e) the number of edges in the sentence graph was high; and (f) there were relatively few key lemmas that also occurred in the course textbook index. This pattern would arise in essays with high

average pair-wise sentence similarity but with low variation in word adjacency. We interpret this factor as reflecting the students' phrase structure creativity.

Factor 2 explained 13.4% of the variance in the data set. Essays scored highly on this factor if (a) the key lemmas in the short version of the assignment question occurred frequently in the essay compared to other essays; (b) the bigrams in the long version of the assignment question occurred frequently in the essay; (c) many key lemmas in the short version of the assignment question occurred in the essay; (d) the essay had many bigrams; and (e) the key lemmas in the long version of the assignment question occurred frequently in the essay. We interpret this factor as reflecting the students' attention to the terminology in the assignment question.

Factor 3 explained 9.5% of the variance in the data set. Essays scored highly on this factor if (a) there were many paragraphs; (b) there were many headings; (c) there were many true sentences; and (d) the tidied sentences tended to be short. We interpret this factor as reflecting the students' use of fundamental essay components. (Students who used more paragraphs and sentences would have to write shorter sentences to remain within the word limit).

Factor 4 explained 10.7% of the variance in the data set. Essays scored highly on this factor if (a) the number of lemmas was relatively high; (b) the total number of words (including repeats) was relatively high; and (c) the number of tidy words (words after the removal of stop words) was high. We interpret this factor as reflecting established properties of natural language (the average number of inflections per lemma occurring in English prose, and Zipf's law).

Factor 5 explained 7.6% of the variance in the data set. Essays scored highly on this factor if (a) a high proportion of the essay's true sentences occurred in the introduction; and (b) many of the top 30 key sentences occurred in the introduction. We interpret this factor as reflecting the quality of the introduction section.

Factor 6 explained 6.4% of the variance in the data set. Essays scored highly on this factor if (a) many of the top 30 key sentences occurred in the conclusion; and (b) a high proportion of the essay's true sentences occurred in the conclusion. We interpret this factor as reflecting the quality of the conclusion section.

Factor 7 explained 8.6% of the variance in the data set. Essays scored highly on this factor if the key lemmas in the textbook index occurred frequently in the essay; and (b) many of the key lemmas in the textbook index occurred in the essay. We interpret this factor as reflecting the students' attention to the terminology in the course textbook.

Finally, the regression method was used to estimate the scores obtained by the 135 essays on each of the seven factors. These factor scores are akin to standard scores (i.e. they have a mean of 0 and a standard deviation of approximately 1).

Regression analysis

A multiple regression analysis was carried out to investigate whether these factor scores predicted the marks that the tutors had awarded the essays. The marks awarded under the first assessment criterion reflected the students' online

contributions and were not directly related to the quality of their essays. Accordingly, the essay marks were adjusted by removing this component, leaving a possible range from 0% to 90%. The actual marks that were awarded to the 135 essays ranged from 24% to 88% with a mean of 63.7%. We noted that the number of references in the reference list had not shown a salient loading on any of the factors (see Table 2). Nevertheless, we considered that it might be important in predicting the overall essay mark, and we therefore included it as a predictor variable.

The overall regression equation was statistically significant, $R^2 = 0.15$, $F(8, 126) = 2.80$, $p = .007$. There was a highly significant effect of the number of references, $B = .36$, $F(1, 126) = 11.78$, $p = .001$, which indicated that the students who cited more references tended to obtain higher marks. More specifically, for citing three extra references students would be expected to achieve an increase of 1 percentage point (i.e. $.36 \times 3$) in their overall mark. There was also a significant relationship with the scores on Factor 1, $B = 2.01$, $F(1, 126) = 4.21$, $p = .04$, which indicated that students who obtained higher scores on this factor also tended to obtain higher marks.

Bearing in mind that most scores on this factor would fall within ± 3 standard deviations of the mean (i.e. between +3 and -3), the students with the highest scores would be expected to obtain marks 12 percentage points (i.e. $2.01 \times [3 - (-3)]$) higher than the students with the lowest scores. None of the other factor scores showed a significant relationship with the students' marks.

Conclusions

Our EssayAnalyser uses state-of-the-art techniques in natural language processing to generate a rich description of the form and content of students' essays. One of the real strengths of the system is that it is domain independent and can be used to analyse any text on any topic. The various attributes that it generates can be reduced to a set of seven relatively independent constructs that explain a high proportion of the variance in the data set. Some of these constructs, especially Factors 4–7, can be explained by properties of mathematics, linguistics or program design. However, Factors 1–3 do not seem to be mere artefacts but reflect important aspects of how students go about writing their essays. Factors 2 and 3 relate to the students' use of terms found in the assignment question and their use of headings to structure their essays. In addition, Factor 1 is a statistically significant predictor of the marks that students' essays receive. This in some measure appears to suggest that linguistic creativity (as reflected in high phrase structure variation and the sparse use of terms found in the course text book, hence a more frequent use of alternative terms *not* found in the course text book) is associated with higher grades.

In the SAFeSEA project, students submit draft essays to a web-based application called OpenEssayist. The essays are processed by EssayAnalyser, and the output from this component is then used by OpenEssayist to derive and present visual representations of the content and structure of the essays. The students can then reflect upon this feedback, particularly with regard to whether the different visual representations capture the intended meaning of their essays (Field et al. 2013; Van Labeke et al. 2013). The prototype system was initially evaluated using genuine essays written by volunteers who had previously taken the Open University course that was described above. The results were used to refine the feedback strategies and the user interface of OpenEssayist. In parallel with this work, we have carried out additional analyses to identify trends and progress markers in students' essay

writing. At the time of writing, the OpenEssayist system is being further evaluated by making it available to the current cohort of students who are taking the same Open University course to evaluate drafts of their assignments before submission.

The SAFeSEA project constitutes an interesting and potentially important application of current techniques in natural language processing. In addition, we believe that an automatic system can provide feedback that is both meaningful and timely and thus support students in the task of drafting their essays. It is therefore likely to enhance the students' experience, their engagement with academic studies and their learning skills and consequently increase both their retention and their attainment.

Acknowledgements
This work was supported by the Engineering and Physical Sciences Research Council (grant numbers EP/J005959/1 and EP/J005231/1). We are also grateful to the SAFeSEA project coordinator, Bethany Alden, for her assistance.

References

Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python: Analyzing text with the Natural Language Toolkit*. Sebastopol, CA: O'Reilly.

Brin, S., and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30, nos. 1–7: 107–117. http://dx.doi.org/10.1016/S0169-7552(98)00110-X

Chickering, A.W., and Z. F. Gamson. 1987. Seven principles for good practice in undergraduate education. *American Association for Higher Education Bulletin*, March: 3–7. http://www.aahea.org/aahea/articles/sevenprinciples1987.htm

Chickering, A.W., and Z. F. Gamson. 1999. Development and adaptations of the seven principles for good practice in undergraduate education. *New Directions for Teaching and Learning* no. 80: 75–81. http://dx.doi.org/10.1002/tl.8006

Cliff, N. 1988. The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin* 103: 276–279. http://dx.doi.org/10.1037/0033-2909.103.2.276

Comrey, A.L. 1973. *A first course in factor analysis*. New York: Academic Press.

De Winter, J.C.F., D. Dodou, and P.A. Wieringa. 2009. Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research* 44: 147–181. http://dx.doi.org/10.1080/00273170902794206

DiBattista, D., J.O. Mitterer, and G. Leanne. 2004. Acceptance by undergraduates of the immediate feedback assessment technique for multiple-choice testing. *Teaching in Higher Education* 9: 17–28. http://dx.doi.org/10.1080/1356251032000155803

Evans, C. 2013. Making sense of assessment feedback in higher education. *Review of Educational Research* 83: 80–120. http://dx.doi.org/10.3102/0034654312474350

Field, D., S. Pulman, N. Van Labeke, D. Whitelock, and J.T.E. Richardson. 2013.

Did I really mean that? Applying automatic summarisation techniques to formative feedback. Paper presented at the 9th international conference on Recent Advances in Natural Language Processing, September 7–13, in Hissar, Bulgaria. http://lml.bas.bg/ranlp2013/docs/RANLP_main.pdf

Freeman, L. 1977. A set of measures of centrality based on betweenness. *Sociometry* 40: 35–41. http://dx.doi.org/10.2307/3033543

Gibbs, G., and C. Simpson. 2004–05. Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education*, no. 1: 3–31. http://insight.glos.ac.uk/tli/resources/lathe/Documents/issue%201/articles/simpson.pdf

Hattie, J., and H. Timperley. 2007. The power of feedback. *Review of Educational Research* 77: 81–112. http://dx.doi.org/10.3102/003465430298487

Mihalcea, R., and P. Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, ed. D. Lin and D. Wu, 404–411. Stroudsburg, PA: Association for Computational Linguistics. http://aclweb.org/anthology/W/W04/W04-3252.pdf

Nelson, M.M., and C.D. Schunn. 2009. The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science* 37: 375–401. http://dx.doi.org/10.1007/s11251-008-9053-x

Newman, M.E.J. 2008. Mathematics of networks. In *The new Palgrave dictionary of economics*, 2nd ed., vol. 5, ed. S.N. Durlauf and L.E. Blume, 465–470. Houndmills: Palgrave Macmillan.

Nicol, D.J., and D. Macfarlane-Dick. 2006. Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education* 31: 199–218. http://dx.doi.org/10.1080/03075070600572090

O'Connor, B.P. 2000. SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, and Computers* 32: 396–402. http://dx.doi.org/10.3758/BF03200807

Pitcher, N., J. Goldfinch, and C. Beevers. 2002. Aspects of computer-based assessment in mathematics. *Active Learning in Higher Education* 3: 159–176. http://dx.doi.org/10.1177/1469787402003002005

Richardson, J.T.E. 1978. A factor analysis of self-reported handedness. *Neuropsychologia* 16: 747–748. http://dx.doi.org/10.1016/0028-3932(78)90010-6

Sapnas, K.G., and R.A. Zeller. 2002. Minimizing sample size when using exploratory factor analysis for measurement. *Journal of Nursing Measurement* 10: 135–154. http://dx.doi.org/10.1891/jnum.10.2.135.52552

Seale, J.K. 2006. *E-learning and disability in higher education: Accessibility research and practice*. London: Routledge.

Van Labeke, N., D. Whitelock, D. Field, S. Pulman, and J. Richardson. 2013. What is my essay really saying? Using extractive summarization to motivate reflection and redrafting. Paper presented at the Artificial Intelligence in Education Workshop on Formative Feedback in Interactive Learning Environments, July 13, in Memphis, TN. http://ceur-ws.org/Vol-1009/0810.pdf

Walker, D.J., K. Topping, and S. Rodrigues. 2008. Student reflections on formative e-assessment: Expectations and perceptions. *Learning, Media and Technology* 33: 221–234. http://dx.doi.org/10.1080/17439880802324178

Whitelock, D., and A. Brasher. 2006. Developing a roadmap for e-assessment: Which way now? In *10th CAA International Computer Assisted Assessment Conference*, ed. M. Danson, 487–501. Loughborough: Loughborough University. http://caaconference.co.uk/pastConferences/2006/proceedings/Whitelock_D_Brasher_A_v2.pdf

Whitelock, D., and Y. Raw. 2003. Taking an electronic mathematics examination from home: What the students think. In *Computer based learning in science: Vol. 1. New technologies and their applications in education*, ed. C.P. Constantinou and Z.C. Zacharia, 701–713. Nicosia: University of Cyprus, Department of Educational Sciences. http://cblis.uniza.sk/cblis-cd-old/2003/3.PartB/Papers/Math_Ed/Whitelock.pdf