Developing and testing a self-assessment and tutoring system

Øistein E. Andersen
iLexIR
Streets, 62 Hills Road
Cambridge, CB2 1LA
and@ilexir.co.uk

Helen Yannakoudakis
Cambridge English
1 Hills Road
Cambridge, CB1 2EU
yannakoudakis.h@cambridgeenglish.org

Fiona Barker
Cambridge English
1 Hills Road
Cambridge, CB1 2EU
barker.f@cambridgeenglish.org

Tim Parish
iLexIR
Streets, 62 Hills Road
Cambridge, CB2 1LA
tim@ilexir.co.uk

Abstract

Automated feedback on writing may be a useful complement to teacher comments in the process of learning a foreign language. This paper presents a self-assessment and tutoring system which combines an holistic score with detection and correction of frequent errors and furthermore provides a qualitative assessment of each individual sentence, thus making the language learner aware of potentially problematic areas rather than providing a panacea. The system has been tested by learners in a range of educational institutions, and their feedback has guided its development.

1 Introduction

Learning to write a foreign language well requires a considerable amount of practice and appropriate feedback. Good teachers are essential, but their time is limited. As recently shown in a study by Wang et al. (in press) conducted amongst first-year students of English at a Taiwanese university, automated writing evaluation can lead to increased learner autonomy and higher writing accuracy. In this paper, we investigate the merits of a self-assessment and tutoring (SAT) system specifically aimed at intermediate learners of English, at around B2 level in the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001). There are a large number of students at this level, and they should have sufficient knowledge of the language to benefit from the system whilst at the same time committing errors which can be identified reliably.

The system provides automated feedback on learners' writing at three different levels of granularity: an overall assessment of their proficiency, a score for each individual sentence, highlighting well-written passages as well as ones requiring more work, and specific comments on local issues including spelling and word choice.

Computer-based writing tools have been around for a long time, with Criterion (Burstein et al., 2003, which also provides a number of features for teachers) and ESL Assistant (Gamon et al., 2009, not currently available) aimed specifically at second-language learners, but the idea of indicating the relative quality of different parts of a text (sentences in our case) has, to the best of our knowledge, not been implemented previously. This kind of non-specific feedback does not provide a precise diagnosis or immediate cure, but might have the advantage of fostering learning.

In addition to describing the SAT system itself, we present a series of three trials in which learners of English in a number of educational contexts used the system as a tool to work on written responses to specific tasks and improve their writing skills.

2 System

The SAT system is made available to students learning English as a Web service to which they can sign up with a code ('class key') provided by their teacher. Once they have filled in a short demographic questionnaire, the users can respond to one, two, three or more writing tasks. The students can save their work at any time and ask the system to assess the current version of their text, which will

Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pages 32–41,
Atlanta, Georgia, June 13 2013.

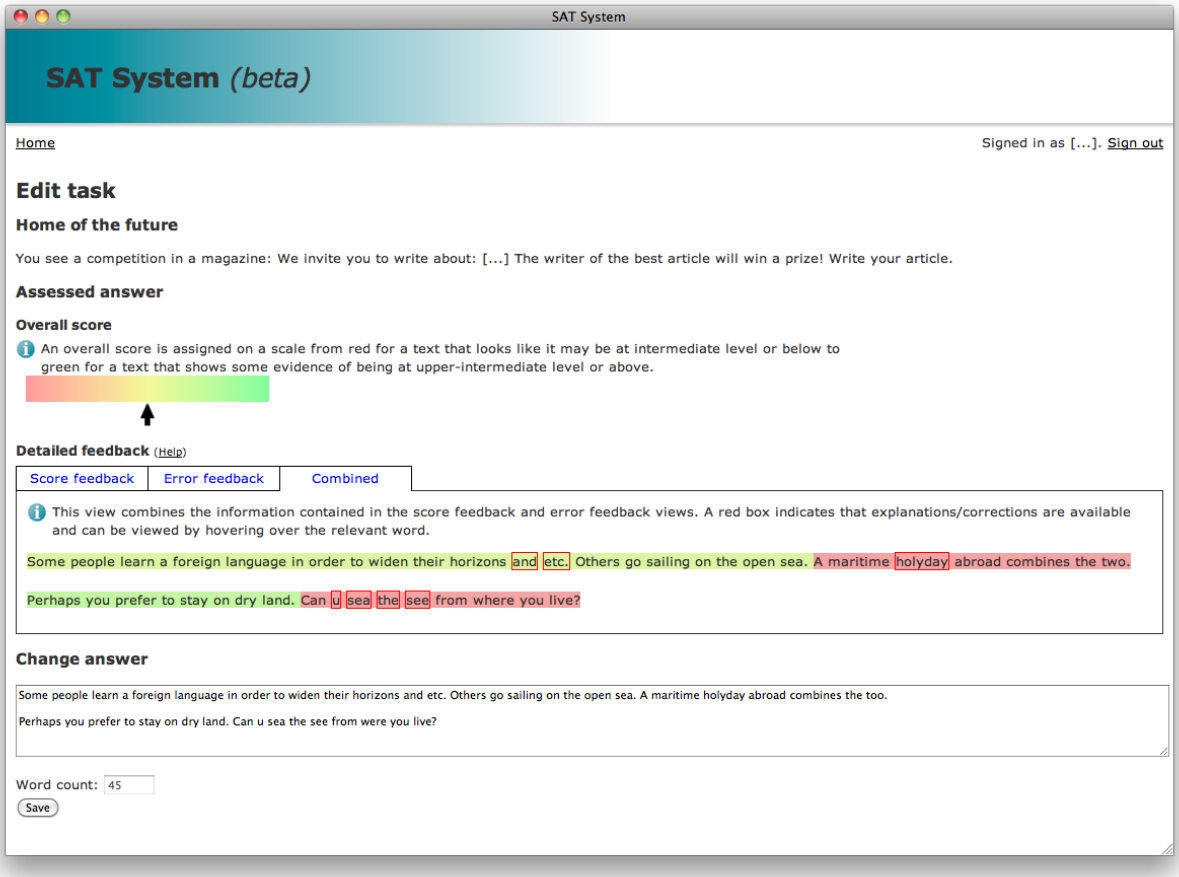

Figure 1: SAT system screen where students can see the automated feedback and revise their piece of writing. The 'score feedback' and 'error feedback' views are shown in Figures 2 and 3.

give feedback as shown in Figure 1 and described in more detail in the following subsections. Assessment times are currently around 15sec, which facilitates incremental and exploratory editing of a text to improve it, giving the students the ability to try out different ways of correcting a problematic turn of phrase. The teacher can see which students have signed up and look at the last saved version of their responses. Finally, the students are asked to answer a few questions about their experience with the system.

2.1 Text assessment

The SAT system provides an overall assessment of someone's proficiency by automatically analysing and scoring the text as a whole. There is a large body of literature with regard to automated text scoring systems (Page, 1968; Rudner and Liang, 2002; Attali and Burstein, 2006; Briscoe et al., 2010). Existing systems, overviews of which have been published in various studies (Dikli, 2006; Williamson, 2009; Shermis and Hamner, 2012), involve a large range of techniques, such as discriminative and generative machine learning, clustering algorithms and vectorial semantics, as well as syntactic parsers.

We approach automated text assessment as a supervised machine learning problem, which enables us to take advantage of existing annotated data. We use the publically-available First Certificate in English (FCE) dataset of upper-intermediate learner English (Yannakoudakis et al., 2011) and focus on assessing general linguistic competence. Systems that measure English competence directly are easier and faster to deploy, since they are more likely to be reusable and generalise better across different genres than topic-specific ones, which are not immediately

usable when new tasks are added, since the model cannot be applied until a substantial amount of manually annotated responses have been collected for a specific prompt.

Following previous research, we employ discriminative ranking, which has been shown to achieve state-of-the-art results on the task of assessing free-text writing competence (Yannakoudakis et al., 2011). The underlying idea is that high-scoring texts (or 'scripts') should receive a higher rank than low-scoring ones. We train a linear ranking perceptron (Bös and Opper, 1998) on features derived from previous work (namely, lexical and grammatical properties of text) and compare it to our previous model (Yannakoudakis et al., 2011), which is trained using ranking Support Vector Machines (Joachims, 2002). Our new perceptron model achieves 0.740 and 0.765 Pearson product-moment (r) and Spearman's rank correlation coefficient (ρ) respectively between the gold and predicted scores; this is comparable to our previous SVM model, which achieves 0.741 and 0.773, and the differences are not significant.

In order to provide scoring feedback[1] based on the predictions of our model, we use visual presentations. Visualisation techniques allow us to go beyond the mere display of a number, can stimulate the learners' visual perceptions, and, when used appropriately, information can be displayed in an intuitive and easily interpretable way. Furthermore, aesthetics in computer-based interfaces have been shown to have an effect on the users. For example, Ben-Bassat et al. (2006) have found an interdependence between perceived aesthetics and usability in questionnaire-based assessments, and have shown that users' preferences are not necessarily based only upon performance; aesthetics also play a role.

More specifically, we assign an overall score on a scale from red for a text that looks like it may be at intermediate level or below to green for a text that shows some evidence of being at upper-intermediate level (the level assessed by the FCE exam) or above (*i.e.,* advanced). This is illustrated in Figure 1 below the *Overall score* section, where an arrow is used to indicate the level of text quality on a colour gradient defined by the two extreme points, red and green. A text with the highest score possible would indicate that the learner has potentially shown evidence of being at a level higher than that assessed by FCE, the latter, of course, being dependent on the extent to which higher-order linguistic skills are elicited by the prompts. On the contrary, a very low score indicates poor linguistic abilities corresponding to a lower level.

Although exams that encompass the full range of language proficiency exhibited at different stages of learning are hard to design, the FCE exam, benchmarked at the B2 level and reserving some of its score range for performances beneath and beyond, allows us to roughly estimate someone's proficiency as being far below, just below, around or above an upper intermediate level. The task of predicting attainment levels has recently started to receive attention (Dickinson et al., 2012; Hawkins and Filipović, 2012).

2.2 Sentence evaluation

The second component of the SAT system automatically assesses and scores the quality of individual sentences, independently of their context. The challenge of assessing intra-sentential quality lies in the limited linguistic evidence that can be extracted automatically from relatively short sentences for them to be assessed reliably, in addition to the difficulty in acquiring annotated data, since rating a response sentence by sentence is not something examiners typically do and would therefore require an additional and expensive manual annotation effort.

Previous work has primarily focused on automatic content scoring of short answers, ranging from a few words to a few sentences (Pulman and Sukkarieh, 2005; Attali et al., 2008; Mohler et al., 2011; Ziai et al., 2012). On the other hand, scoring of individual sentences with respect to their linguistic quality, specifically in learner texts, has received considerably less attention. Higgins et al. (2004) devised guidelines for the manual annotation of sentences in learner texts, and evaluated a rule-based approach that classifies sentences with respect to clarity of expression based on grammar, mechanics and word usage errors; however, their system performs binary classification, whereas we are focusing on scoring sentences. Writing instruction tools, such as Criterion (Burstein et al., 2003), give advice on stylistic

[1] Note that ranks can be transformed to scores through linear regression, while correlation remains unaltered as it is invariant to linear transformations.

and organisational issues and automatically detect a variety of errors in the text, though they do not explicitly allow for an overall evaluation of sentences with respect to various writing aspects. The latter, used in combination with an error feedback component (see Section 2.3), can be a useful instrument informing learners about the severity of their mistakes; for example, although sentences may contain some errors, they may still maintain a certain level of acceptability that does not impede communication. Moreover, indicating problematic regions may be better from a pedagogic point of view than detecting and correcting all errors identified in the text.

To date, there is no publically available annotated dataset consisting of sentences marked with a score representing their linguistic quality. Manual annotation is typically expensive and time-consuming, and a certain amount of annotator training is generally required. Instead, we exploit already available annotated data – scores and error annotation in the FCE dataset – and evaluate various approaches, two of which are: a) to use the script-level model (see Section 2.1) to predict sentence quality scores, and b) to use the script-level score divided by the total number of (manually annotated) errors in a sentence as pseudo-gold labels to train a sentence-level model.

As the models above are expected to contain a certain amount of noise, it is imperative that we identify evaluation measures that are indicative of our application – that is, assign higher scores to high-quality sentences compared to low-quality ones – and not only depend on the labels they have been trained on. More specifically, we use correlation with pseudo-gold scores (r_{g} and ρ_{g}; not applicable to the script-level model), correlation with the script-level scores by first averaging predicted sentence-level scores (r_{s} and ρ_{s}), correlation with error counts (r_{e} and ρ_{e}), average precision (AP) and pairwise accuracy. AP is a measure used in information retrieval to evaluate systems that return a ranked list of documents. Herein, sentences are ranked by their predicted scores, precision is calculated at each correct sentence (that is, containing no errors), and averaged over all correct sentences (in other words, we treat sentences with no errors as the 'relevant documents'). Pairwise accuracy is calculated based on the number of times the corrected sentence (available through the error annotation in the FCE dataset) is ranked higher than the original one written by the candidate, ignoring sentences without errors. Correlation with error counts, average precision and pairwise accuracy are particularly important as they reflect more directly the extent to which good and bad sentences are discriminated. Again, in both cases, we employ a linear ranking perceptron.

We conducted a series of experiments on a separate development set to evaluate the performance of features beyond the ones used in the script-level model. The final results, reported in Table 1, are calculated on the FCE test set (Yannakoudakis et al., 2011).

Our best configuration is model b, which achieves the highest results according to most evaluation measures with a feature space consisting of 1) error counts identified through the absence of word trigrams in a large background corpus, 2) phrase-structure rules, 3) presence of frequent errors, as well as the number of words defining an error, as described in Section 2.3, 4) the presence of main verbs, nouns, adjectives, subordinating conjuctions and adverbs, 5) affixes and 6) the presence of clausal subjects and modifiers. The texts were parsed using RASP (Briscoe et al., 2006).

Model a, the script-level model, does not work as well at the sentence level. However, it does perform better when evaluated against script-level scores (r_{s} and ρ_{s}), and this is expected given that it is trained directly on gold script-level scores. On the other hand, this evaluation measure is not as indicative of good performance in our application as the others, as it does not take into account the varying quality of individual sentences within a script.

Training the script-level model with different feature sets (including those utilised in the sentence-level model) did not yield an improvement in performance (the results are omitted due to space restrictions). Additional experiments were conducted to investigate the effect of training the sentence-level model with different pseudo-gold labels (*e.g.*, additive/subtractive pseudo-gold scores rather than divisive/multiplicative), but the results are not reported here as the difference in performance was not substantial.

Table 1 shows that better performance can be achieved with our pseudo-gold labels, used to train a model at the sentence level, rather than gold la-

	Model a	Model b
r_g	—	0.550
ρ_g	—	0.646
r_s	0.572	0.385
ρ_s	0.578	0.301
r_e	–0.111	–0.750
ρ_e	–0.078	–0.702
AP	0.393	0.747
Pairwise		
Correct	0.608	0.703
Incorrect	0.359	0.204

Table 1: Results on the FCE test set for the script-level model (a) and our model (b).

bels at the script level. To evaluate this further, we trained a sentence-level model using the script-level scores as labels (that is, sentences within the same script are all assigned the same label/score). However, this did not improve performance (again, the results are omitted due to space restrictions). We also point out that the best-performing feature space (described above) is based on text properties that are more likely to be present in relatively short sentences (*e.g.,* the presence of main verbs), compared to those used for script-level models in previous work (Yannakoudakis et al., 2011), such as word and part-of-speech bigrams and trigrams, which may be too sparse for a sentence-level model.

Analogously to what we did to present the overall score, we developed a sentence score feedback view to indicate the general quality of the sentences, as given by our best model, by highlighting each of them with a background colour ranging from green for a well-written sentence, via yellow and orange for a sentence which the system thinks is acceptable, to dark orange and red for a sentence which may have a few problems. Figure 2 shows how the SAT system evaluates and colour-codes a few authentic student-written sentences containing errors, as well as their corrected counterparts based on the error-coding in the FCE test set. Overall, the system correctly identifies correct and incorrect versions of each sentence, attributing a higher score (greener colour) to the corrected sentence in each pair.

2.3 Word-level feedback

Basic spelling checkers have been around since the 1970s and grammar checkers since the 1980s (Kukich, 1992), but misleading 'corrections' may be bewildering (Galletta et al., 2005), and the systems do not always focus on the kinds of error frequently committed, even less so in the case of learners as was pointed out early on by Liou (1992), who tested commercial grammar checkers on and developed a system for detecting common errors in Taiwanese learners' writing.

For word-level feedback within the SAT system, we have implemented a method similar to one we have used earlier in the context of pre-annotation of learner corpora (Andersen, 2011). To ensure high precision and good coverage of local errors typically committed by learners, error rules are generated from the Cambridge Learner Corpus (CLC) (Nicholls, 2003) to detect word unigrams, bigrams and trigrams which have been annotated as incorrect at least five times and at least ninety per cent of the times they occur. This way, rules can be extracted from the existing error annotation in the corpus, obviating the need for manually constructed mal-rules, although the rules obtained by the two different methods may to some extent be complementary. In addition to corpus-derived rules, many classes of incorrect but plausible derivational and inflectional morphology are detected by means of rules derived from a machine-readable dictionary. Many mistakes are still not detected, but precision has been found to be more important in terms of learning effect (Nagata and Nakatani, 2010), and errors missed by this module will often give lower sentence scores.

Figure 3 illustrates some types of error detected by the system. The feedback text is generated from a small number of templates corresponding to different categories of error marked up in the CLC.

We are currently working on extending this part of the system with more general rules in addition to word n-grams, *e.g.,* part-of-speech tags and grammatical relations, in order to detect more errors without loss in precision.

3 Trials

After the SAT system had been developed, a series of trials were set up in order to test the online sys-

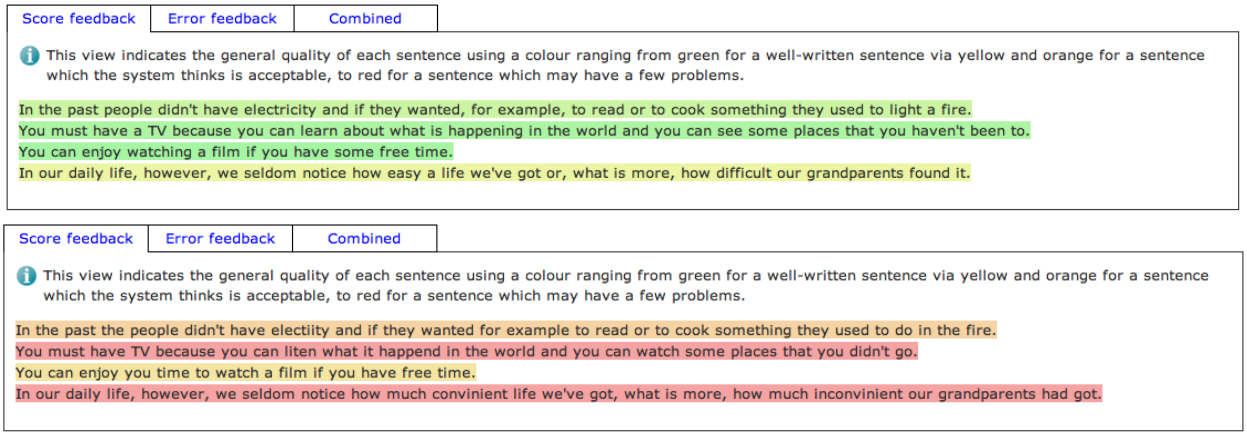

Figure 2: Examples of correct sentences (top) and incorrect ones (bottom) colour-coded by the SAT system.

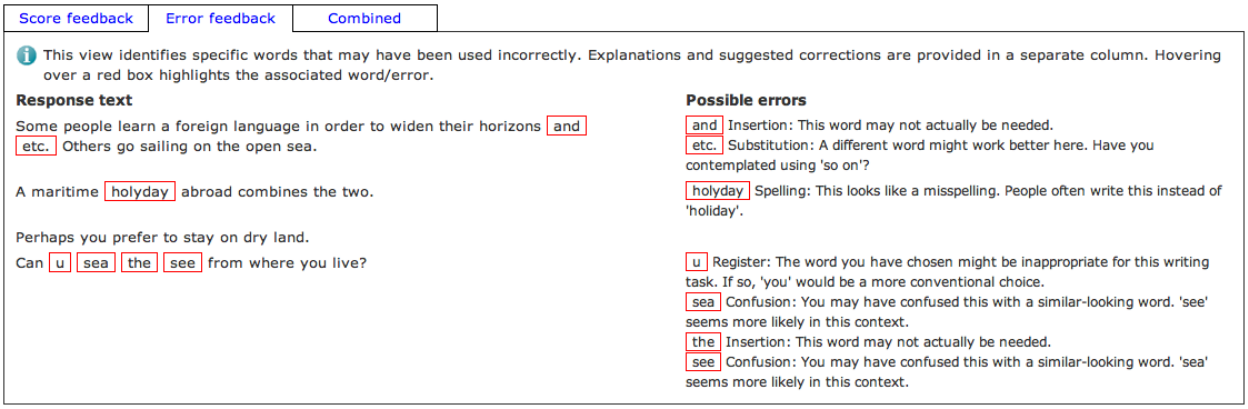

Figure 3: The *error feedback* view identifies specific words that may have been used incorrectly. Explanations and suggested corrections are provided in a separate column. The system actually proposes two different corrections for *and etc.*, namely *etc.* and *and so on*; the user will have to choose one or the other. The confusion between the verb *see* and the noun *sea* is identified, but the *the* is not actually unnecessary; in this case, the system has been led astray by the surrounding errors.

tem and to collect feedback from language learners and their teachers in a variety of contexts. Three trials were undertaken in November 2012, December 2012 and in March 2013, with changes made to the system between each pair of trials.

English Profile Network member institutions were contacted who had access to language learners and who had previously participated in data collection for the English Profile Programme[2]. Teachers at universities, secondary schools and private language schools signed up for two or more trials so that their learners could use and provide feedback on several iterations of the SAT system. Certificates of participation were offered to encourage involvement in the trials.

Ten institutions were involved from nine countries, namely Belgium, the Czech Republic, France, Lithuania, Poland, Romania, Russia, Slovakia and Spain. Eight universities, one secondary school and one private language school were represented, including specialist and generalist institutions of educational sciences, agricultural science, veterinary medicine and foreign languages. Each trial had between 4 and 8 institutions taking part, and each institution participated in two or three trials with many students undertaking more than one trial.

All students who took part in the trials, over 450

[2]See *www.englishprofile.org*

in total, were expected to be at or above the upper-intermediate (CEFR B2) level as this was the level at which the SAT system was designed to function.

Three initial sets of tasks were developed for the planned system trials, each set consisting of three short written prompts which asked the users to write on a specified topic for a particular purpose, for example:

> *Daily life*
>
> *Your English class is going to make a short video about daily life in your town. Write a report for your teacher, suggesting which activities should be filmed, and why.*

Tasks were based on retired questions from an international proficiency test at B2 level of the CEFR. Each task was given a short name which was shown in the SAT system in order for the users to select the most interesting or relevant task for themselves.

A short set of instructions was produced for both teachers and students which was emailed to the main contact in each institution and passed on to their colleagues, teachers and students who were interested in taking part in the trial.

The trials operated as follows:

- The main institutional contact receives an invitation to participate in the trials.
- Interested institutions receive instructions and confirm the number of class keys required (sign-up codes for the system).
- Main contact and teachers at each institution log in and work through the system as if they are a language learner, by completing a demographic questionnaire, writing 1–3 tasks which are assessed by the system, and finally completing a short user satisfaction questionnaire.
- Students work through the SAT system either with the support of their teacher in class or remotely.

3.1 SAT system usage

During Trial 1, on the busiest day there were 155 submissions and the highest number of users on a single day was 32. These figures indicate that all users were submitting their work for assessment more than once, which suggests that the system is being used in an iterative fashion as envisaged. During Trial 2, the busiest day saw more than twice as many submissions as during the first trial (442), and the most people online on any one day almost doubled to 62. Across both trials we collected around 3000 submissions in total, including revisions; the average number of revisions for a submitted piece of writing is 3.2 with the highest figure being 54 revisions (see Table 2 for details). This suggests that some users write their first response, then make changes to one word or phrase at a time, resulting in such a large number of revisions. When more than one revision has been submitted, the score given by the system to the last revision is higher than that given to the initial revision in over 80% of the cases. Current changes to the system allowing system administrators to check on intermediate versions of submitted texts are underway.

Revisions	Count
1	292
2	272
3	142
4	78
5	50
6	28
7	15
8	25
9	11
10	14
11–15	21
16–20	6
20–	5

Table 2: Number of revisions per task response.

3.2 Feedback

In addition to looking at the writing submitted by users of the system, there was both numerical and written feedback available to the system developers. This was used to suggest changes to the system at subsequent trials.

As can be seen from Table 3, user satisfaction scores were generally high and increased from Trial 1 to Trial 2. In the first pilot, the written feedback from instructors was generally positive whilst

	Trial 1	Trial 2
Using the SAT system helps me to write better in English.	3.80	3.92
I find the SAT system useful for understanding my mistakes.	3.74	3.96
I think the sentence colouring is useful.	3.74	4.15
I think the word-level information [error feedback] is useful.	3.86	4.12
The SAT system is easy to use.	4.45	4.49
The feedback on my writing is clear.	3.80	3.93
If you have used the SAT system before, has it improved since the last time?		3.86

Table 3: Average feedback scores on a scale from 1 (strongly disagree) to 5 (strongly agree).

the learner feedback was mixed, especially when it comes to sentence evaluation:

> *In summary, I liked this system, because the sentence colouring suggests me to think about my writing style, mistakes, what I should improve, change. This system is not like a teacher, who checks all our errors, but makes us develop our critical thinking, which is the most important for writing especially. [...]*

> *It's okay the way of colouring system, the problem is that it doesn't tell you specifically what's wrong with constructions so you have think what you failed.*

The fact that the system provides almost immediate feedback has been appreciated:

> *I like that the paragraphs which I wrote assesed so quickly. ... Secondly, I really like that student can correct his text till it gets ideal.*

Users have also made suggestions for improvements, which have been essential for deciding which parts of the system should be developed further.

3.3 System changes

As a result of feedback and the team's extensive use of the system, after each trial changes were made both to the on-screen experience and behind the scenes. After Trial 1, the system was amended to enable users to see paragraph breaks in the corrected version (which before had not been shown in the assessed view of the text). There was also a new error view with permanently visible explanations and examples and an additional question on the feedback questionnaire which asked whether users felt the system had improved since the previous time they used it. Behind the scenes, the server was upgraded to cope with anticipated demand and code was written so that administrators could review statistics on usage.

Words	Count
0– 99	540
100–199	1,294
200–299	928
300–399	201
400–499	67
500–999	26
1,000–	36

Table 4: Number of words per submission.

At the time of writing the third SAT system trial was underway. In the first two trials the total number of words collected was over 600,000 with an average response length of around 1100 characters or 200 words. Encouragingly, there were many longer responses including twelve over 1080 words in length and the longest written to date is 1773 words. These figures indicate that the system is not restrictive, but encourages and inspires students to write. Table 4 gives an overview of the script length distribution.

Following two successful trials, the third trial aimed to involve new and existing users and to provide more detailed teacher feedback.

4 Conclusions

In this paper, we described a tool that provides feedback to learners of English at three different levels of granularity: an overall assessment of their proficiency, assessment of individual sentences, and diagnostic feedback on local issues including spelling and word choice. We argued that the use of visual-

isation techniques is important, as they allow us to go beyond the mere display of a number, can stimulate the learners' visual perceptions, and can display information in an intuitive and easily interpretable way. The usefulness and usability of the tool as a whole, as well as of its components, was confirmed through questionnaire-based evaluations, where, for example, the perceived usefulness of the sentence colouring received an average of 4.15 on a 5-point scale.

The first component of the SAT system, script-level assessment, uses a machine learner to predict a score for a text and roughly estimate someone's proficiency level based on lexical and grammatical features. The second component allows for an automatic evaluation of the linguistic quality of individual sentences. We proposed a method for generating sentence-level scores, which we use for training our model. Using this method, we were able to learn what features can be used to evaluate linguistic quality of (relatively short) sentences. Indicating problematic regions via highlighting of sentences may be better from a pedagogic point of view than detecting and correcting all errors identified in the text. The third component automatically provides diagnostic feedback on local errors with high precision on the basis of a few templates, without relying on manually crafted rules.

The trials undertaken so far have improved the functionality of the system in regard to what is on offer to teachers and their students, but they have also provided the basis for further research and development to enhance the system's functionality and design and move towards wider deployment. We plan to continue improving the methodologies used for providing feedback to learners, as well as adding further functionality, such as L1-specific feedback. Another logical next step would be to continue towards lower levels of granularity, moving from the sentence as the unit of assessment to clauses and phrases, which may be particularly beneficial for more advanced language users who write longer and more complex sentences.

Acknowledgements

Special thanks to Ted Briscoe and Marek Rei, as well as to the anonymous reviewers, for their valuable contributions at various stages.

References

Øistein E. Andersen. 2011. Semi-automatic ESOL error annotation. *English Profile Journal*, 2.

Yigal Attali and Jill Burstein. 2006. Automated essay scoring with e-Rater v.2.0. *Journal of Technology, Learning, and Assessment*, 4(3):1–30.

Yigal Attali, Don Powers, Marshall Freedman, Marissa Harrison, and Susan Obetz. 2008. Automated Scoring of short-answer open-ended GRE subject test items. Technical Report 04, ETS.

Tamar Ben-Bassat, Joachim Meyer, and Noam Tractinsky. 2006. Economic and subjective measures of the perceived value of aesthetics and usability. *ACM Transactions on Computer-Human Interaction*, 13(2):210–234.

Siegfried Bös and Manfred Opper. 1998. Dynamics of batch training in a perceptron. *Journal of Physics A: Mathematical and General*, 31(21):4835–4850.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *ACL-Coling'06 Interactive Presentation Session*, pages 77–80.

Ted Briscoe, Ben Medlock, and Øistein E. Andersen. 2010. Automated assessment of ESOL free text examinations. Technical Report UCAM-CL-TR-790, University of Cambridge, Computer Laboratory.

Jill Burstein, Martin Chodorow, and Claudia Leacock. 2003. Criterion: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the fifteenth annual conference on innovative applications of artificial intelligence*, pages 3–10.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Markus Dickinson, Sandra Kübler, and Anthony Meyer. 2012. Predicting learner levels for online exercises of Hebrew. In *Proceedings of the Seventh Workshop on Innovative Use of NLP for Building Educational Applications*, pages 95–104. Association for Computational Linguistics.

Semire Dikli. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment*, 5(1).

Dennis F. Galletta, Alexandra Durcikova, Andrea Everard, and Brian M. Jones. 2005. Does spell-checking software need a warning label? *Communications of the ACM*, 48(7):82–86.

Michael Gamon, Claudia Leacock, Chris Brockett, William B Dolan, Jianfeng Gao, Dmitriy Belenko, and

Alexandre Klementiev. 2009. Using statistical techniques and web search to correct ESL errors. *Calico Journal*, 26(3):491–511.

John A. Hawkins and Luna Filipović. 2012. *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*. English Profile Studies. Cambridge University Press.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining*, pages 133–142.

Karen Kukich. 1992. Techniques for automatically correcting words in text. *ACM Computing Surveys*, 24(4):377–439.

Hsien-Chin Liou. 1992. An automatic text-analysis project for EFL writing revision. *System: The International Journal of Educational Technology and Language Learning Systems*, 20(4):481–492.

Michael A.G. Mohler, Razvan Bunescu, and Rada Mihalcea. 2011. Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Ryo Nagata and Kazuhide Nakatani. 2010. Evaluating performance of grammatical error detection to maximize learning effect. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 894–900, Stroudsburg, PA, USA. Association for Computational Linguistics.

Diane Nicholls. 2003. The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In Dawn Archer, Paul Rayson, Andrew Wilson, and Tony McEnery, editors, *Proceedings of the Corpus Linguistics conference*, volume 16 of *Technical Papers*, pages 572–581. University Centre For Computer Corpus Research on Lanugage, Lancaster University, Lancaster.

Ellis B. Page. 1968. The use of the computer in analyzing student essays. *International Review of Education*, 14(2):210–225.

Stephen G. Pulman and Jana Z. Sukkarieh. 2005. Automatic short answer marking. In *Proceedings of the second workshop on Building Educational Applications Using natural language processing*, pages 9–16.

Lawrence M. Rudner and Tahung Liang. 2002. Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2):3–21.

Mark D. Shermis and Ben Hamner. 2012. Contrasting state-of-the-art automated scoring of essays: analysis. Technical report, The University of Akron and Kaggle.

Ying-Jian Wang, Hui-Fang Shang, and Paul Briody. In press. Exploring the impact of using automated writing evaluation in English as a foreign language university students' writing. *Computer Assisted Language Learning*.

David M. Williamson. 2009. A framework for implementing automated scoring. In *Proceedings of the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education*, San Diego, CA.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the workshop on Building Educational Applications Using natural language processing*, pages 190–200.