Using the E-assessment Approach to Provide Diagnostic Feedback: the Case for Automated Assessment as Evidence of Learners’ Progress
Dr Ardeshir Geranpayeh
Head of Automated Assessment & Learning

29 September 2016
The learning ecology

Connecting 4 worlds of learning
The learning ecology

- Learning
- Interaction
- Observation
- Feedback
- Tasks
The learning ecology

- Learning
  - Tasks
  - Feedback
  - Observation
  - Interaction
The learning ecology
LOA in practice

- Learning objectives
- Course
- Task

Common European Framework of Reference (CEFR):
- C2: Proficient user
- C1
- B2: Independent user
- B1
- A2
- A1
LOA in practice

Common European Framework of Reference (CEFR):
- C2: Proficient user
- C1: Independent user
- B2: Basic user
- B1
- A2
- A1

Learning objectives → course → Task → Language activity
LOA in practice

Learning objectives → Course → Task → Language activity

Common European Framework of Reference (CEFR): C2, C1, B2, B1, A2, A1

Teacher observation
LOA in practice

learning objectives

Common European Framework of Reference (CEFR)

- C2 Proficient user
- C1
- B2 Independent user
- B1
- A2 Basic user
- A1

Course

Task

Language activity

Teacher observation

Interpretation

#EATPconf
LOA in practice
LOA in practice

Teacher decision-making

Learning objectives

Course

Task

Language activity

Teacher observation

Interpretation

Informal record

Common European Framework of Reference (CEFR)

A2

Basic user

A1

Independent user

B1

Proficient user

C1

C2

Record

#EATPconf
LOA in practice

Learning objectives

Task

Language activity

Teacher observation

Feedback & modify learning objectives

Teacher decision-making

Record

Informal record

Common European Framework of Reference (CEFR)

- C2: Proficient user
- C1
- B2: Independent user
- B1
- A2
- A1: Basic user

#EATPconf
LOA in practice

Common European Framework of Reference (CEFR):
- A1: Basic user
- B1: Independent user
- B2: Proficient user
- C1: Lower intermediate
- C2: Upper intermediate

Learning objectives

Course

Task

Language activity

Feedback & modify learning objectives

Teacher decision-making

Record

Informal record

Teacher observation

Interpretation

#EATPconf
LOA in practice

Learning objectives

Course

Task

Language activity

Feedback & modify learning objectives

Teacher decision-making

Record

Structured record

Informal record

Teacher observation

Interpretation

Record of achievement

Common European Framework of Reference (CEFR)
- C2: Proficient user
- C1
- B2: Independent user
- B1
- A2
- A1: Basic user

#EATPconf
LOA in practice

Learning objectives

Common European Framework of Reference (CEFR)

- C2 Proficient user
- C1
- B2 Independent user
- B1
- A2 Basic user
- A1

Teacher decision-making

Monitoring of performance

Feedback & modify learning objectives

Task

Language activity

External exam

Teacher observation

Record

Informal record

Structured record

Record of achievement

Interpretation

Monitor performance
Figure 1: Model showing macro and micro levels in the LOA cycle

- **Macro level (setting and monitoring targets)**
  - Learning objectives (high-level, detailed)
  - CB placement testing
  - Summative monitoring
  - Frame of reference (CEFR)
  - Record of achievement

- **Micro level (materials, classroom practice)**
  - Task
  - Feedback, modify learning objectives
  - Teacher observation
  - Language activity
  - Record
  - Structured record
  - Informal record

- **Adaptivity**
  - Teacher decision making
  - Interpretation

- **Input Analysis**
  - Record
  - Informal record

- **Digital tech-mediated tasks**
  - CB assessment including auto rating

- **Digital data**
  - Digital tech-mediated learning activities

- **Data analysis**
  - Digital data

- **Individualised student model**
  - Language theory building, description
Digital technology ...

- Plays a crucial role in facilitating LOA:
  - by extending learning beyond the physical classroom
  - by enabling new forms of learning interaction
  - by capturing new forms of evidence for learning

- It assists the teacher:
  - by scaffolding the learners' use of language in more authentic contexts of use
  - by removing the administrative burden of collecting and processing large amounts of information
The task: automated writing feedback

- Automated writing feedback
  - Automatically evaluate the quality of writing and provide immediate feedback
- Challenges
  - Provide accurate, effective and detailed feedback
  - Provide pedagogically useful feedback like human teachers
Deployment

Advantages

- Prompt detailed feedback
- Promote writing development
- Facilitate self-assessment and self-tutoring
- Application of constant assessment criteria
- Reduced workload
- Cost-effective approach to teaching / grading
Script-level feedback

- Text Assessment: Overall assessment of someone's proficiency by scoring the text as a whole

1. Assess general linguistic competence
   a) Gather annotated data
   b) Identify textual features considered to be proxies for intrinsic qualities of writing competence
   c) Predict score using weighted combination of features (Machine Learning)
   d) Evaluate predicted scores

2. Provide scoring feedback
Deployment
Script-level feedback: Feature Space

- Word Sequences
  - believe (unigram)
  - suggest idea (bigram)
  - the people is (trigram)

- Part of Speech sequences
  - VV0 VV0 (e.g., keep develop)
  - NN2 VVG (e.g., children smiling)

- Grammatical constructions
  - V1/modal bse/+ (e.g., can only travel in July)
  - S/pp-ap s-r (e.g., for better or worse, he left)
  - T/txt-frag (e.g., but know Kim knew)

- Error rate & other features (text, length, complexity …)
Script-level feedback: Evaluation

<table>
<thead>
<tr>
<th>Features</th>
<th>Pearson's correlation $r$</th>
<th>Spearman's correlation $\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>word seq</td>
<td>0.601</td>
<td>0.598</td>
</tr>
<tr>
<td>+PoS seq</td>
<td>0.682</td>
<td>0.687</td>
</tr>
<tr>
<td>+text length</td>
<td>0.692</td>
<td>0.689</td>
</tr>
<tr>
<td>+syntax</td>
<td>0.714</td>
<td>0.712</td>
</tr>
<tr>
<td>+error rate</td>
<td>0.741</td>
<td>0.773</td>
</tr>
<tr>
<td>Upper bound</td>
<td>0.796</td>
<td>0.792</td>
</tr>
</tbody>
</table>
Comparison with previous work
Regression Vs Ranking

<table>
<thead>
<tr>
<th>Model</th>
<th>Pearson's correlation r</th>
<th>Spearman's correlation ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.720</td>
<td>0.750</td>
</tr>
<tr>
<td>Our Model (Ranking)</td>
<td>0.750</td>
<td>0.785</td>
</tr>
</tbody>
</table>
Script-level feedback: Evaluation

Overall score

An overall score is assigned on a scale from red to green. Red is for text that looks like it may be at CEFR level B1 or below. Green is for text that shows some evidence of being at CEFR level B2 or above. Once you have had a look at your feedback, try to use it to improve your writing, and then submit it again.

Now, improve your answer

Some people learn a foreign language in order to widen their horizons and etc.

Perhaps you prefer to stay on dry land. Can you see the sea from were you live?

Word count: 31
Websites A

Write your text in English using the task you've chosen. You can enter your text directly, or paste in work you've already done. If you want to save your writing but aren't ready to get it assessed, click on Save. You can come back and review it when you like by clicking on Edit. When you are ready to have your writing assessed, click on Save and Submit. Once you've got your feedback use it to try and improve your writing – then click Save and Submit. Remember the tasks are there to help you think of what to write. You won't be marked on your ability to complete the tasks, just your writing.

Describe two websites that you use. What do you like about them?

Feedback

Overall score

An overall score is assigned on a scale from red to green. Red is for text that looks like it may be at CEFR level B1 or below. Green is for text that shows some evidence of being at CEFR level B2 or above. Once you have had a look at your feedback, try to use it to improve your writing, and then submit it again.

Detailed feedback

Combined feedback allows you to see the information contained in the Sentence feedback and Error feedback together on one page. A red box indicates that explanations or corrections are available and can be viewed by hovering over the word. An orange box indicates words that might need attention to improve your results, but for which the system doesn't have a suggestion.

Some people learn a foreign language in order to widen their horizons and use the computer to find websites of places to go on holiday. Others go sailing on the open sea. A maritime holiday abroad combines the two. Perhaps you prefer to stay on dry land. Can you see the sea from where you live?
### Word-level feedback: error detection & correction

<table>
<thead>
<tr>
<th>Trigrams</th>
<th>Error</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>he] want [to</td>
<td>AGV</td>
<td>wants</td>
</tr>
<tr>
<td>to] thanks [all</td>
<td>FV</td>
<td>thank</td>
</tr>
<tr>
<td>are] to [old</td>
<td>SX</td>
<td>too</td>
</tr>
<tr>
<td>’s] interesting [place</td>
<td>MD</td>
<td>an+</td>
</tr>
<tr>
<td>is] need [to</td>
<td>MD</td>
<td>a+</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Bigrams</th>
<th>Error</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>of] whole</td>
<td>MD</td>
<td>the+</td>
</tr>
<tr>
<td>This [why</td>
<td>MV</td>
<td>+ is</td>
</tr>
<tr>
<td>few] absence</td>
<td>AGN</td>
<td>absences</td>
</tr>
<tr>
<td>listening] at</td>
<td>RT</td>
<td>to</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unigrams</th>
<th>Error</th>
<th>Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>beloveds</td>
<td>C</td>
<td>beloved</td>
</tr>
<tr>
<td>disappointment</td>
<td>S</td>
<td>disappointment</td>
</tr>
<tr>
<td>singed</td>
<td>IV</td>
<td>sang</td>
</tr>
</tbody>
</table>
Word-level feedback: error detection & correction

Response text

Some people learn a foreign language in order to widen their horizons and etc. Perhaps you prefer to stay on dry land.

Can you see the see from were you live?

Possible errors

and Insertion: This word may not be needed.

e等 Substitution: A different word might be better here. Perhaps ‘so on’ is better.

sea Confusion: Is this the right word? Did you mean to write ‘see’?

the Insertion: This word may not be needed.

see Confusion: Is this the right word? Did you mean to write ‘sea’?
Sentence-level feedback

• Sentence evaluation
  • Assess and score the quality of individual sentences, independently of their context

• Challenges
  • Limited linguistic evidence that can be extracted automatically
  • Difficulty in acquiring annotated data
Model 2: sentence-level model with pseudo-gold labels: \( \frac{\text{score}}{\text{errors}} \)

### Feature set

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Main verbs, nouns, adjectives, subordinating conjunctions and adverbs</td>
</tr>
<tr>
<td>2</td>
<td>Clausal subjects and modifiers</td>
</tr>
<tr>
<td>3</td>
<td>Affixes</td>
</tr>
<tr>
<td>4</td>
<td>Phrase-structure rules</td>
</tr>
<tr>
<td>5</td>
<td>Errors</td>
</tr>
<tr>
<td>6</td>
<td>Number of words forming an error</td>
</tr>
</tbody>
</table>
In the past people didn't have electricity and if they wanted, for example, to read or to cook something they used to light a fire.

You must have a TV because you can learn about what is happening in the world and you can see some places that you haven't been to.

You can enjoy watching a film if you have some free time.

In our daily life, however, we seldom notice how easy a life we've got or, what is more, how difficult our grandparents found it.

In the past the people didn't have electricity and if they wanted for example to read or to cook something they used to do in the fire.

You must have TV because you can listen what it happened in the world and you can watch some places that you didn't go.

You can enjoy your time to watch a film if you have free time.

In our daily life, however, we seldom notice how much convenient life we've got, what is more, how much inconvenient our grandparents had got.
Feedback

Overall score

An overall score is assigned on a scale from red to green. Red is for text that looks like it may be at CEFR level B1 or below. Green is for text that shows some evidence of being at CEFR level B2 or above. Once you have had a look at your feedback, try to use it to improve your writing, and then submit it again.

Detailed feedback (Help)

Combined feedback allows you to see the information contained in the Sentence feedback and Error feedback together on one page. A red box indicates that explanations or corrections are available and can be viewed by hovering over the word. An orange box indicates words that might need attention to improve your results, but for which the system doesn’t have a suggestion.

Some people learn a foreign language in order to widen their horizons and etc. Perhaps you prefer to stay on dry land.

Can you see the see from were you live?

Now, improve your answer

Some people learn a foreign language in order to widen their horizons and etc. Perhaps you prefer to stay on dry land.

Can you see the see from were you live?

Word count: 31
User Satisfaction

<table>
<thead>
<tr>
<th>Statement</th>
<th>Trial 1</th>
<th>Trial 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Using W&amp;I helps me to write better in English</td>
<td>3.80</td>
<td>3.92</td>
</tr>
<tr>
<td>I find W&amp;I useful for understanding my mistakes</td>
<td>3.74</td>
<td>3.96</td>
</tr>
<tr>
<td>I think the sentence colouring is useful</td>
<td>3.74</td>
<td>4.15</td>
</tr>
<tr>
<td>I think the word-level information [error feedback] is useful</td>
<td>3.86</td>
<td>4.12</td>
</tr>
<tr>
<td>W&amp;I is easy to use</td>
<td>4.45</td>
<td>4.49</td>
</tr>
<tr>
<td>The feedback on my writing is clear</td>
<td>3.80</td>
<td>3.93</td>
</tr>
<tr>
<td>If you have used W&amp;I before, has it improved since the last time?</td>
<td>—</td>
<td>3.86</td>
</tr>
</tbody>
</table>

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

- User-driven development between trials

#EATPconf
Latest live figures

• 18,000 New Submissions
• 102,000 submissions, including revisions
  ▪ 2/3rd of initial submissions revised
  ▪ 50% more than once
  ▪ 11% more than 10 times
• Average number of revisions: 6.7
• Max number of revisions: 2164
Future work

• Improve methodologies used for providing error feedback
• Add further functionality
  • L1-specific feedback
  • Discourse organisation feedback
  • Task achievement feedback
Acknowledgments

Jones, N. and N. Saville (2016)
How does it work (link to Video)
Thank you for listening

Geranpayeh.a@CambridgeEnglish.org