Crowdsourcing
Second Language learner data:

experiences and prospects

Elena Volodina
University of Gothenburg, Språkbanken, Sweden
The Rise of Crowdsourcing

Remember outsourcing? Sending jobs to India and China is so 2003. The new pool of cheap labor: everyday people using their spare cycles to create content, solve problems and even do corporate R & D.

Jeff Howe, 2006, Wired magazine
WELCOME TO 
THE AGE OF THE CROWD

Use “plugged-in enthusiasts”,
“take advantage of the networked world”,
“discover ways to tap the latent talent of the crowd”

Jeff Howe
ICALL tools for Second language (L2) learning

Target group

Language skill

Resources

Tools & algorithms

Prototypes

Evaluation
Curious “time & effort” fact:

Data vs experiments
Lark Trills for Language Drills
Text-to-speech technology for language learners

- Dictation and spelling exercise
- Focus on
  - evaluation of the quality of TTS
  - finding ways to give feedback on spelling errors
Lark Trills for Language Drills
Text-to-speech technology for language learners

- Dictation and spelling exercise
- Focus on
  - evaluation of the quality of TTS
  - finding ways to give feedback on spelling errors
### Fully automatic

- self-study mode
- test mode
- timed test
- word
- inflected word
- phrase
- sentence
- performance

Generate

repeat slower faster

### Result Tracker

<table>
<thead>
<tr>
<th>Exercise name</th>
<th>Correct/Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learners/spelling-word, self-study</td>
<td>1/2</td>
</tr>
</tbody>
</table>

### Train spelling, word level

Type the word you hear

<table>
<thead>
<tr>
<th>Nr</th>
<th>Word</th>
<th>Correct answer</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>matematik</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Submit

matematik

- Saldo morphology: matematik
- Wikipedia: matematik
- Wiktionary: matematik
- Monica: listen to pronunciation
User input (all attempts) logged to a database

Pipeline for word & (inflected word) levels
For each correct item (base form + word class) we store:
- session ID (no personal data, such as L1)
- incorrect spelling(s)
L2 spelling error database, SPEED

- `<LexicalEntry uid="LexicalEntry-58d3459f-5acb-43f8-b60e-deb45a986c56">
  <Sense id="speed--kelly-6950" uid="Sense-b1d45016-bdb5-4584-9ce5-11f780ecbf8a'/>
  <word lang="swe" pos="AV" uid="word-4d9ed4cb-83b7-4293-a786-ff11f398e2d2">förträfflig</word>
  <misspelling sessionID="2013-05-13-22-27-28" time="22:58:25" uid="misspelling-3ba31d83-f1ad-4c99-bc33-2c6a3a3c7849">förträvlig</misspelling>
- `<modification uid="modification-4673c2b2-63a1-4c3c-b2a5-c6a80bc5dd20">
  <feat att="updatedBy" val="lærka" uid="feat-ec50eb25-d870-4f53-b86c-ed620e3a332c'/>
  <feat att="modificationDateTime" val="2013-05-13T22:58:26.01+02:00" uid="feat-4032acab-afa3-42c3-b6b0-3f904e345b76'/>
  <feat att="modificationAccepted" val="pending" uid="feat-2e43c1b4-1106-4364-8df6-3991a4da6578'/>
  <feat att="modificationComment" val="" uid="feat-22d76dd2-c26f-4ccd-b7a9-e6997082e0cd'/>
</modification>
</LexicalEntry>`
## Error data

<table>
<thead>
<tr>
<th>Error types</th>
<th>Nr,%</th>
<th>Example (correct → <em>incorrect</em>)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Competence-based errors</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Consonant doubling</td>
<td>28</td>
<td><em>stoppa → stopa</em></td>
</tr>
<tr>
<td>Diacritics (å, ä, ö)</td>
<td>23</td>
<td><em>högre → hogre</em></td>
</tr>
<tr>
<td>Phonetic errors (e.g. voiced vs voiceless)</td>
<td>25</td>
<td><em>relevans → relevanz</em></td>
</tr>
<tr>
<td>Consonant clusters (phoneme-grapheme mappings, incl. cases of homonyms)</td>
<td>20</td>
<td><em>skön → sjön</em></td>
</tr>
<tr>
<td>Other (unclassifiable)</td>
<td>4</td>
<td><em>Israel → visträv</em></td>
</tr>
<tr>
<td>Performance-based errors</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>Typos (neighbouring keys, addition, deletion, insertion, replacement)</td>
<td>17</td>
<td><em>förbättra → förb'ttra</em></td>
</tr>
<tr>
<td>Across one word (phrases &amp; sentences)</td>
<td>28</td>
<td><em>se en bild → sen bild</em></td>
</tr>
</tbody>
</table>
Spelling error?

- no
  - Feedback action (FA): Mark as correct

- yes
  - Is it a real word?
    - no
      - Tip 1: Unknown word. Please, check it again
    - yes
      - Does the error occur with characters with diacritics? (å, ä, ö)
        - no
          - Tip 2: Pay attention to the specific Swedish letters (å, ä, ö)
        - yes
          - Tip 2: Pay attention to the consonant doubling
            - no
              - Error analysis
            - yes
              - Tip 2: Pay attention to alternative spellings of the sound [xx]: x1, x2...
SPEED
SPElling Error Database

Advantages of collecting a corpus by applying this method: participants are quickly attracted, while cost, time and effort of collecting a corpus are reduced.

THIS is RESEARCH DATA!
And we need more of it!
What is infrastructure?

"'Infrastructure'? — You mean like rocks and sticks?"
An electronic research infrastructure

- (free accessible) data in electronic format
- technical platform for exploring data, including tools and algorithms for data analysis, and visualization
- a set of tools and technical solutions for new data collection and preparation, including data processing and annotation
- a network of experts in the relevant disciplines, incl. legal and ethical questions
Partners

- University of Gothenburg: NLP, L2, language assessment
  *Elena Volodina, Julia Prentice, Monica Reichenberg*

- Stockholm university: NLP, L2
  *Mats Wirén, Gunlög Sundberg*

- Uppsala university: NLP
  *Beata Megyesi*

- Umeå university: L2/assessment
  *Lena Granstedt*
Guess what?

- Riksbankens Jubileumsfond, infrastructure project IN16-0464:1
- 2017-2019
SweLL: electronic research infrastructure on Swedish learner language

- SweLL – Swedish Learner Language
- Lärka-based L2 infrastructure
  ... as a unit under Språkbanken's infrastructure
  ... in the context of CLARIN
Our focus is on...

- L2 essays (writing)
- exercise logs (reading and listening comprehension, vocabulary and grammar training)
- NO speech data – yet
- target group: adult learners
L2 “alternative” data

- Logs – acc. to a defined research interest
- Steps:
  - Implement an activity for learners
  - Prepare database for storing (structured) data
  - Implement a way to browse logs, visualize statistics etc
  - If necessary – add extra annotation steps (manual, automatic)
Pilot 1 on L2 “alternative” data

- Identifying most predictive features for a language proficiency level (for diagnostic purposes)
  - Multi-word expressions
  - Syntactic properties (e.g. word order)
  - Knowledge of word morphology (e.g. inflections)

David Alfter
L2 “alternative” data (logs)

Exercise type evaluation

Bundled gaps (variant 1)
Which word fits into these gaps? Each gap contains the same word. Write the word.

Hennes _______ var på hans lår, gned in värme i hans kalla ben.
En annan taxi tar _______ om skolbarnen.
Novelty hade flera trumf på _______.
I första _______ har hon spelat dragspel och fiol.

Evaluation
For which levels is this exercise type relevant?

[ ] A1  [ ] A2  [ ] B1  [ ] B2  [ ] C1

Comments

https://sprakbanken.gu.se/larkalabb/exeeval
Pilot 2 on L2 “alternative” data

- Automatic assigning new words to a proficiency level
  - We predict the level automatically
  - Learners (of a known level) get the word in an exercise (or a series of exercises)
  - We see whether learners can cope with it
L2 “alternative” data (logs)

https://spraakbanken.gu.se/larkalabb/wordguess

Word guess

Tries: 0/7

Definition:
rör sig tyst och försiktigt för att inte bli upptäckt, tassar (pres ind aktiv)

Score: 0

Hjälp:
slip; steal, sneak, creep
The ultimate goal

L2 infrastructure activity development cycle

1. **Update resources, tools, algorithms**
   - NLP: input data analysis, user modeling

2. **Resources, tools, algorithms**

3. **Research scenarios**
   - Activity prototype
   - Activity generation

4. **Reuse**
   - Annotation of answers (assessors/researchers)
   - Activity in use

5. **Publish activity**
   - Logging answers to database(s)
   - L2 expert analysis
   - L2 expert approval
   - GUI for assessors/teachers (assessor view)

6. **GUI for researchers (researcher view)**

7. **GUI for students (student view)**
Where will this lead?
Thank you!

Questions?