SWELLEX, PRODUCTIVE L2 LEARNER VOCABULARY AND MORE

David Alftr
University of Gothenburg
It all started with... a demo

https://sprakbanken.gu.se/larka/texteval
Texteval

- Text analysis platform
- Assessment of learner written and expert written texts
Texteval

- Machine learning and readability measures
  - ML: predict overall level of text
  - Readability measures
    - Number of sentences
    - Number of tokens
    - Average sentence/token length
    - LIX score
Texteval

- Word-level CEFR highlighting
  - Depends on graded word lists
  - Out-of-vocabulary words
Word lists

- **SVALex** (François et al., 2016)
  - *COCTAILL corpus* (Volodina et al., 2014)
  - *Receptive knowledge*

- **SweLLLex** (Volodina et al., 2016)
  - *SweLL corpus* (Volodina et al., 2016)
  - *Productive knowledge*

- **Kelly list** (Volodina and Kokkinakis, 2012)
  - *L1 web corpus*
Word list format (SVALex & SweLLLex)

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS-tag</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>bil</td>
<td>NN_UTR</td>
<td>430.218</td>
<td>1234.207</td>
<td>728.9847</td>
<td>422.283</td>
<td>363.5446</td>
<td>618.8567</td>
</tr>
<tr>
<td>överge</td>
<td>VB</td>
<td>0</td>
<td>0</td>
<td>7.3203</td>
<td>24.5182</td>
<td>39.6516</td>
<td>17.2695</td>
</tr>
<tr>
<td>rättvisa</td>
<td>NN_UTR</td>
<td>0</td>
<td>0</td>
<td>3.6601</td>
<td>25.6189</td>
<td>26.4344</td>
<td>13.6602</td>
</tr>
<tr>
<td>kilo</td>
<td>NN_NEU</td>
<td>0</td>
<td>302.0833</td>
<td>145.1229</td>
<td>65.0611</td>
<td>13.2172</td>
<td>89.8907</td>
</tr>
<tr>
<td>resa</td>
<td>VB</td>
<td>166.300</td>
<td>375.2582</td>
<td>450.3526</td>
<td>298.4905</td>
<td>330.4297</td>
<td>356.362</td>
</tr>
<tr>
<td>låg</td>
<td>JJ</td>
<td>0</td>
<td>49.315</td>
<td>125.922</td>
<td>217.3103</td>
<td>252.1311</td>
<td>156.126</td>
</tr>
<tr>
<td>så klart</td>
<td>ABM_MWE</td>
<td>0</td>
<td>16.2635</td>
<td>81.6019</td>
<td>45.5033</td>
<td>13.2172</td>
<td>38.1738</td>
</tr>
</tbody>
</table>
## Word list format (Kelly list)

<table>
<thead>
<tr>
<th>ID</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Freq</td>
<td>2624.032</td>
</tr>
<tr>
<td>Word per Million</td>
<td>23017.26</td>
</tr>
<tr>
<td>CEFR level</td>
<td>A1</td>
</tr>
<tr>
<td>Source</td>
<td>SweWaC</td>
</tr>
<tr>
<td>Grammar marker</td>
<td>att</td>
</tr>
<tr>
<td>Item</td>
<td>vara (vardagl. va)</td>
</tr>
<tr>
<td>POS</td>
<td>verb</td>
</tr>
<tr>
<td>Example</td>
<td>e.g. var så god!</td>
</tr>
</tbody>
</table>
**Word list format (Kelly list)**

<table>
<thead>
<tr>
<th>ID</th>
<th>88</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Freq</td>
<td>2624.032</td>
</tr>
<tr>
<td>Word per Million</td>
<td>23017.26</td>
</tr>
<tr>
<td>CEFR level</td>
<td>A1</td>
</tr>
<tr>
<td>Source</td>
<td>SweWaC</td>
</tr>
<tr>
<td>Grammar marker</td>
<td>att</td>
</tr>
<tr>
<td>Item</td>
<td>vara (vardagl. va)</td>
</tr>
<tr>
<td>POS</td>
<td>verb</td>
</tr>
<tr>
<td>Example</td>
<td>e.g. var så god!</td>
</tr>
</tbody>
</table>
Mapping from distributions to labels

- Different possible approaches:
  - First occurrence (Gala et al., 2013; Gala et al., 2014)
  - Maximum
  - Thresholding (Alfter et al., 2016)
  - ...
Mapping distributions to levels

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS-tag</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>bil</td>
<td>NN_UTR</td>
<td>430.218</td>
<td>1234.207</td>
<td>728.9847</td>
<td>422.283</td>
<td>363.5446</td>
<td>618.8567</td>
</tr>
<tr>
<td>överge</td>
<td>VB</td>
<td>0</td>
<td>0</td>
<td>7.3203</td>
<td>24.5182</td>
<td>39.6516</td>
<td>17.2695</td>
</tr>
<tr>
<td>rättvisa</td>
<td>NN_UTR</td>
<td>0</td>
<td>0</td>
<td>3.6601</td>
<td>25.6189</td>
<td>26.4344</td>
<td>13.6602</td>
</tr>
<tr>
<td>kilo</td>
<td>NN_NEU</td>
<td>0</td>
<td>302.0833</td>
<td>145.1229</td>
<td>65.0611</td>
<td>13.2172</td>
<td>89.8907</td>
</tr>
<tr>
<td>resa</td>
<td>VB</td>
<td>166.300</td>
<td>375.2582</td>
<td>450.3526</td>
<td>298.4905</td>
<td>330.4297</td>
<td>356.362</td>
</tr>
<tr>
<td>låg</td>
<td>JJ</td>
<td>0</td>
<td>49.315</td>
<td>125.922</td>
<td>217.3103</td>
<td>252.1311</td>
<td>156.126</td>
</tr>
<tr>
<td>så klart</td>
<td>ABM_MWE</td>
<td>0</td>
<td>16.2635</td>
<td>81.6019</td>
<td>45.5033</td>
<td>13.2172</td>
<td>38.1738</td>
</tr>
</tbody>
</table>
First occurrence: A1
Maximum: B1
Threshold: A2
### Mapping distributions to levels

<table>
<thead>
<tr>
<th>Lemma</th>
<th>POS-tag</th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>bil</td>
<td>NN_UTR</td>
<td>430.218</td>
<td>1234.207</td>
<td>728.9847</td>
<td>422.283</td>
<td>363.5446</td>
<td>618.8567</td>
</tr>
<tr>
<td>överge</td>
<td>VB</td>
<td>0</td>
<td>0</td>
<td>7.3203</td>
<td>24.5182</td>
<td>39.6516</td>
<td>17.2695</td>
</tr>
<tr>
<td>rättvisa</td>
<td>NN_UTR</td>
<td>0</td>
<td>0</td>
<td>3.6601</td>
<td>25.6189</td>
<td>26.4344</td>
<td>13.6602</td>
</tr>
<tr>
<td>kilo</td>
<td>NN_NEU</td>
<td>0</td>
<td>302.0833</td>
<td>145.1229</td>
<td>65.0611</td>
<td>13.2172</td>
<td>89.8907</td>
</tr>
<tr>
<td>resa</td>
<td>VB</td>
<td>166.300</td>
<td>375.2582</td>
<td>450.3526</td>
<td>298.4905</td>
<td>330.4297</td>
<td>356.362</td>
</tr>
<tr>
<td>låg</td>
<td>JJ</td>
<td>0</td>
<td>49.315</td>
<td>125.922</td>
<td>217.3103</td>
<td>252.1311</td>
<td>156.126</td>
</tr>
<tr>
<td>så klart</td>
<td>ABM_MWE</td>
<td>0</td>
<td>16.2635</td>
<td>81.6019</td>
<td>45.5033</td>
<td>13.2172</td>
<td>38.1738</td>
</tr>
</tbody>
</table>
First occurrence: A2
Maximum: C1
Threshold: B2
Sparse data

- Word lists created from sparse data
Japan (PM)

C1?
Trädgårdsästare (NN)
Out-of-vocabulary words

- katt (cat)
Single-word lexical complexity

- Word length
- Syllables
- Suffix length
- Gender
- Homonymy
- Polysemy
- Compounds
- N-grams
Feature extraction

- Input to machine learning algorithm
- Different algorithms tested
  - SVM
  - Logistic regression
  - MLP
- Different combinations of features tested
Topic distributions

- CEFR proficiency levels correspond roughly to topics
  - A1: introductions, greetings
  - A2: personal life, family
  - B1: school, leisure, personal interests

(Council of Europe, 2001, p. 26)
Topic distributions

- Extract topic lists from
  - COCTAILL corpus: 33 topics
  - Swedish FrameNet (SweFN++): 1010 topics

- Retain only most predictive words per list (TF-IDF)

- Add topic distribution to feature vector
Results SweLLlex

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>102</td>
<td>37</td>
<td>21</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>A2</td>
<td>41</td>
<td>147</td>
<td>58</td>
<td>33</td>
<td>28</td>
</tr>
<tr>
<td>B1</td>
<td>19</td>
<td>56</td>
<td>157</td>
<td>80</td>
<td>37</td>
</tr>
<tr>
<td>B2</td>
<td>9</td>
<td>46</td>
<td>71</td>
<td>120</td>
<td>32</td>
</tr>
<tr>
<td>C1</td>
<td>8</td>
<td>26</td>
<td>44</td>
<td>58</td>
<td>207</td>
</tr>
</tbody>
</table>
Results SweLLex

- Accuracy: 50%
- F1 score: 0.51
## Results SVALex

<table>
<thead>
<tr>
<th></th>
<th>A1</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>244</td>
<td>64</td>
<td>26</td>
<td>15</td>
<td>10</td>
</tr>
<tr>
<td>A2</td>
<td>34</td>
<td>342</td>
<td>174</td>
<td>121</td>
<td>42</td>
</tr>
<tr>
<td>B1</td>
<td>9</td>
<td>134</td>
<td>474</td>
<td>277</td>
<td>73</td>
</tr>
<tr>
<td>B2</td>
<td>7</td>
<td>71</td>
<td>197</td>
<td>1774</td>
<td>172</td>
</tr>
<tr>
<td>C1</td>
<td>2</td>
<td>25</td>
<td>67</td>
<td>231</td>
<td>1037</td>
</tr>
</tbody>
</table>
Results SVALex

- Accuracy: 69%
- F1 score: 0.65
Siwoco

- Automatic prediction of single word lexical complexity
- MLP classifier

- Demo

https://spraakbanken.gu.se/larkalabb/siwoco
Validation through crowdsourcing

- Idea: use crowd sourcing to validate correctness of target level assignment
- Learner proficiency model predicts current learner level
- Check response time and accuracy for items of different levels
  - Items of learner level: Faster response time and higher accuracy
  - Items of (predicted) learner level with lower response time/lower accuracy possibly higher level
Resource creation

- Self-updating and self-validating graded vocabulary list
  - New unseen words graded by algorithm
  - Untested words given to learners
  - Validation/estimation of level through crowdsourcing
  - Improvement of the resource
Questions? Comments?
david.alfter@gu.se