Automatic assessment of second language proficiency

Klára Vaníčková
Original plan

- Longitudinal study of students’ progress in Czech as a foreign language
- Collect data every week/two weeks in preparatory courses (from 0 to B2 in 9 months)
- Complications
  - COVID-19 unclear situation about incoming students
  - Financing annotators + respondents – not found feasible system
    - failed GAUK proposal
New task

- Categorize texts *for/by* students of Czech as a foreign language according their proclaimed language proficiency level by complexity and readability measures
- Language proficiency – defined in *Common European Framework of Reference for Languages*
- Solving by TOEFL, EVALD…
Motivation

- Level assessment tool for purpose of exams
  - Certified exams
  - Exam for permanent residency
  - Exam for citizenship
- Self-assessment tool
- Scientific pedagogical tool (selection of appropriate reading texts)
Realization

DATA for STUDENTS
- textbooks
- blogs, adaptive reading
- prescription documents

DATA by STUDENTS
- MERLIN
- CZESL-sgt
  - new data from UJOP B2 - C1
  - new longitudinal data

INTERCORP for validating features

FEATURE EXTRACTION
(CTAP - Common Text Analysis Platform)

MACHINE LEARNING

LEVEL of PROFICIENCY
TEXT for STUDENTS

LEVEL of PROFICIENCY
TEXT by STUDENT

EVALUATING FEATURES
for CZECH
Common Text Analysis Platform (CTAP)

- Collected around 600 complexity features: EN, DE, IT, FR, NL, PT, CS
- Plain text -> NLP "annotators" -> rule-based feature "extractors" -> analysis -> visualization
- Infrastructure ready, just plug in CS NLP, write "annotators", re-use (adjust?) extractors available, add own extractors (?)
- User-friendly web GUI
Adaptation of international complexity formulas to Czech

**RE** = 206.835 – (1.015 x ASL) – (84.6 x ASW)

**RE** = Readability Ease

**ASL** = Average Sentence Length (i.e., the number of words divided by the number of sentences)

**ASW** = Average number of syllables per word (i.e., the number of syllables divided by the number of words)

<table>
<thead>
<tr>
<th>text</th>
<th>Flesch RE</th>
</tr>
</thead>
<tbody>
<tr>
<td>jsem-blizky-cloveku-ve-vykonu-trestu</td>
<td>25</td>
</tr>
<tr>
<td>jsem-v-exekuci</td>
<td>6</td>
</tr>
<tr>
<td>prava-a-povinnosti-dluznika/</td>
<td>32</td>
</tr>
<tr>
<td>/web/msp/podanistiznosti-k-eslp</td>
<td>1</td>
</tr>
<tr>
<td>ministerstvo-spravedlnosti</td>
<td>-29</td>
</tr>
</tbody>
</table>

**very simple texts for laymen!**

90-100 : Very Easy
80-89 : Easy
70-79 : Fairly Easy
60-69 : Standard
50-59 : Fairly Difficult
30-49 : Difficult
0-29 : Very Confusing

**rather clear and concise press release**
Thank you!