Revita: a System for Language Learning and Supporting Endangered Languages

Anisia Katinskaia, Javad Nouri, Roman Yangarber

University of Helsinki

22 May 2017

Introduction

- Computer Science Department, University of Helsinki
- Research group in Computational Linguistics
  - Academy Project: a computational tool for language learning
  - ... for supporting less-resourced, endangered languages

Revita project

Specific aims:

- Tools to help user improve competence in language
- $\rightarrow$ not aimed at beginners: intermediate/advanced levels
- endangered languages:
  - Erzya
  - Meadow Mari
  - Udmurt
  - Komi-Zyrian
  - North Saami
  - Sakha

Other uses emerged:

- Swedish, Finnish, Russian.
- Other languages can be added.

Main ideas and features

- Stimulate *active* learning by providing:
  - exercises for active language production
  - possibility to upload stories interesting to the user
  - $\rightarrow$ unlimited supply of exercises, because they are generated automatically
  - $\rightarrow$ exercises based on learners' answers given so far
- Grammar and vocabulary practice
- Multiple-choice and 'cloze' quizzes, crosswords, and flashcards.

Main practice mode

- Choose a text to practice
- Present pieces of text (snippets) to the student in order
- Several words in the snippet will be chosen for quizzes
- Multiple-choice quiz: selected words are presented with distractors
- 'Cloze' quiz: selected words are removed, base forms (lemmas) are presented as hints
- Student's task: guess the correct form of the word—based on lemma and context of the story

Example of exercise

"Topelius kertoo Maamme kirja eri maakunnista"

(*"Topelius tells book Our Land about the different provinces."*)

"Topelius kertoo Maamme kirjassaan eri maakunnista"

- user will receive immediate feedback about answers
- correct form = form found in the story

Crosswords, translation and flashcards

- Student can get a translation of any word in the current snippet
- All words that the student has clicked to get translations are saved to flashcards

Crosswords:

- 40-50 words from story are used to generate a crossword
- user can request additional hint for missing words, which are their grammatical base forms (lemmas)

Generating exercises

Pipeline of generating an exercise from a loaded story:

- tokenize text, extract title, analyze by morphological analyzer;
- extract base forms, parts of speech, grammatical tags from analyses;
- extract from text all words and combinations of words that can serve as candidates for exercises;

Ambiguous words

A surface form is considered ambiguous if it has more than one different lemma. Russian form “жил” has two morphological bases:

- “жить” (live-INF, *“to live”*)
- “жила” (sinew-NOM.SG, *“sinew”*).

In both cases forms “жил” have two different analyses:

- “жил” (live-PST.MASC.SG, *“he lived”*)
- “жил” (sinew-GEN.PL, *“sinew”*).

Word combinations as candidates for exercises

Combinations of words are chosen by Revita based on language-specific rules.

- [pos=adj, case=X, number=Y, gender=Z] [pos=noun, case=X, number=Y, gender=Z];
- [word=в, pos=prep] [pos=noun, case=loc or acc].

Examples of combinations derived by these rules:

- "красивой девушке"
  beautiful-Fem.Dat.Sg girl-Fem.Dat.Sg
  "... [to] a beautiful girl (dative)"
- "в доме"
  in house-M.Loc.Sg
  "In a/the house"

Choosing candidates

Revita uses history to compute weights for exercise candidates.

- Examples always answered correctly by the learner receive low probability;
- examples sometimes answered correctly and sometimes incorrectly receive high probability;
- examples that were never answered correctly receive lower weight;
- Revita computes the proximity of the candidates within the snippet
- randomness is applied when choosing from the final set of weighted candidates.

Code-switching disambiguation

Surface form *“пота”* can be Russian or Komi.

- Komi: first-person singular indicative of verb *“потны”* (“to crack”)
- Russian: genitive singular of the noun *“пот”* (“sweat”)

FU-Russian disambiguation

- for all words *w* with both Russian and FU analyses;
- look through the text and check whether *w* has *"friends"*:
  - whether its lemma is equal to the lemma of some *other* surface form *y* in the text.
  - All words without friends are discarded as "risky"
  - If *w* has FU friends in the story, it is highly likely to be a FU word.
- If *w* has friends, examine its *"neighbors"*
  - The word is again discarded as "risky" if it has at least one immediate neighbor with a Russian analysis.

→ The accuracy obtained for Russian and Udmurt was 0.77

Future work

- refine scoring system;
- add the possibility for collaboration to the system;
- assessment of uploaded stories by their difficulty for the learner, and their quality as learning material;
- accepting forms which are not the same as used by the author, but are allowed by the context
- progress assessment, which is important for developing new exercises

Demo ...

revita.cs.helsinki.fi

Thank you!