# Revita: a System for Language Learning and Supporting Endangered Languages

**Anisia Katinskaia, Javad Nouri, Roman Yangarber**
University of Helsinki
Department of Computer Science
first.last@cs.helsinki.fi

## Abstract

We describe a computational system for language learning and supporting endangered languages. The platform provides the user an opportunity to improve her competency through *active* language use. The platform currently works with several endangered Finno-Ugric languages, as well as with Yakut, and Finnish, Swedish, and Russian. This paper describes the current stage of ongoing development.

## 1 Introduction

Revita is an open online platform designed to help support endangered languages, by stimulating *active* language learning. Current focus is on several endangered languages inside the Russian Federation (RF), which have moderate to small numbers of speakers, including several Finno-Ugric (F-U) languages—Udmurt, Meadow Mari, Erzya, Komi-Zyrian, North Saami—and Sakha (Yakut), a Turkic language.[1] The system also works with Finnish, Swedish, and Russian, for several practical reasons. Finnish is structurally very similar to many Uralic languages. Further, texts in many of the target languages often exhibit spontaneous code-switching into Russian, so a Russian component has emerged as an essential feature of the system.

The tool is aimed at people who already possess some competence in the target language—intermediate to advanced students (i.e., not for the very beginners).

The rest of the paper is organized as follows: Section 2 is devoted to a review of prior work in the area of generating "cloze" exercises, Section 3 describes exercise generation in the Revita system and related research problems, and Section 4 presents the conclusions.

## 2 Prior work

Computer-aided language learning (CALL) was first introduced in the 1950s, and since then has developed significantly as technology evolved. We briefly mention some relevant systems, such as PLATO, (Hart, 1981), and (Chapelle and Jamieson, 1983), which was one of the first and most significant systems for teaching and learning languages. Macario was one of the first video programs for learning Spanish (Gale, 1989); the Athena Language-Learning Project (ALLP) combined "interactivity and more primitive drill-and-practice routine" (Murray, 2014); programs like *À la rencontre de Phillippe* (Murray, 2014) allowed learners to act in the learning language environment. Thousands of other programs have been created. Some of the programs, such as Robo-Sensei (Nagata, 2002) and E-Tutor (Heift, 2001), use NLP (natural language processing) techniques, and may be called "intelligent" CALL systems.

Revita's main learning mode involves a type of exercise known as "*cloze*" in the literature, first described in (Taylor, 1953). In a cloze (deletion) test, a portion of text has some of the words removed, and the learner is asked to recover the missing words. Clozes require an understanding of the context, semantics and syntax in order to identify the missing words correctly.

The approach in (Zesch and Melamud, 2014) involves generating *distractors* for vocabulary clozes—multiple-choice questions. The method for generating lists of distractors is as follows. First "context-insensitive inference rules" are used to generate a set of candidate distractors. This set includes the top-N matches for the target word *w* in the corpus—words which share some con-

[1] All F-U languages are inside RF, except Finnish, Hungarian, North Saami, and Estonian.

Anisia Katinskaia, Javad Nouri and Roman Yangarber 2017. Revita: a system for language learning and supporting endangered languages. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa 2017*. Linköping Electronic Conference Proceedings 134: 27–35.

text words with *w*, which harvests words that are in some sense similar. Then the top-M matches are found which appear in exactly the same context as the cloze item ("context-sensitive inference rules"). A distractor blacklist specifies words that should not be used as distractors. In case there are a large number of distractors, ranking is applied to select the most challenging ones. These can be the less frequent distractors in the corpus, or the most similar to the target word (provided that they are not in the blacklist).

Smith et al. (2010) presented an approach to generation of vocabulary clozes, for English only. Their system takes a key (the target word), chooses distractors from a distributional thesaurus, and identifies a collocate that does not occur with the distractors using "Sketch Engine," a corpus query system. Then the system finds a sentence containing the pair. The best sentence should not be long, with sufficient useful context.

Lee and Seneff (2007) describe an approach to generating distractors for learning English prepositions. Distractors are defined in terms of *usability*—only one choice is correct, requiring minimum post-editing time—and in terms of *difficulty* which means that distractors are on the right level of difficulty, neither too wrong nor too challenging, making these choices appropriate for the less proficient language users.

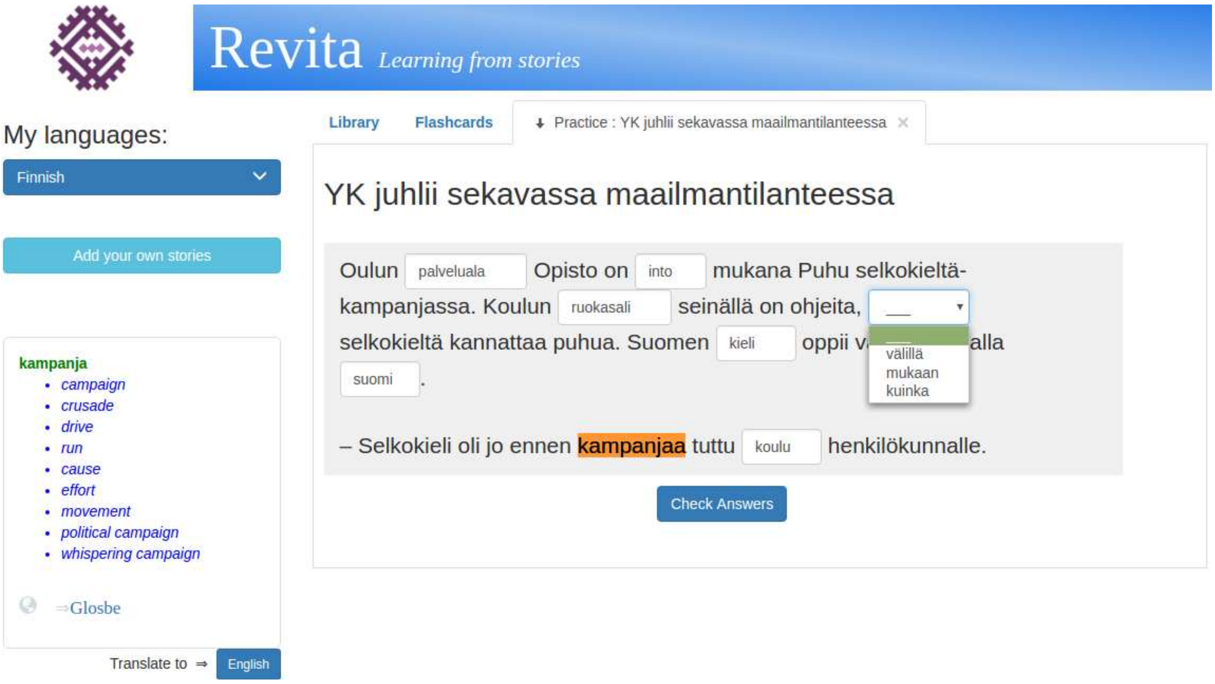
Pino et al. (2008) present a strategy for improving automatically generated cloze and open-cloze (without multiple choice) questions, used by the REAP tutoring system for English as a Second Language vocabulary learning. The system provides the learner with documents retrieved from the Web, filtered for quality and annotated for topic and readability level, to match the student's interest and the model of the student's vocabulary knowledge. For selecting sentences with target words, the system scores sentence complexity, measured by counting the number of clauses, as identified by the Stanford parser. The context of sentences with more clauses is believed to be more well-defined. However, in essence, how well-defined the context is depends on the possibility of replacing the target word with any other word. This can be measured by sum the collocation scores between the target word and other words in the sentence. The authors provide an example: the sentence "I drank a cup of strong (blank) with lemon and sugar" is very well-defined for "tea" because of high collocation scores between "tea" and "strong," "lemon", "sugar", "drink." In absence of these strong collocations, it is less likely to define a target word from the context. This approach showed better results than a baseline.

One of the main problems with this approach is that distractors may fit the context semantically, so open cloze questions can have more than one plausible answer. Also, sentence selection is problematic, since a single sentence may not provide sufficient information for choosing the correct answer.

Brown et al. (2005) present six types of questions for evaluating the level of vocabulary knowledge of REAP system users. This evaluation is used to update the user model of vocabulary knowledge, to provide new texts with 95% of words familiar to the user and 5% of new words. Using WordNet data, the following types of questions were generated: choosing the definition of a word, selecting synonyms and antonyms, hypernym and hyponym question types (completing phrases), and cloze questions. It is shown that there is a correlation between computer-generated questions for assessment of vocabulary skills and human-written questions.

Chen et al. (2006) describe the principles for generation of tests on grammaticality for English language. Tests are based on manually-designed patterns, e.g., the pattern {VB VBG} means that some verb requires a gerund as a complement ("My friends enjoy traveling by plane"). Distractors are usually constructed based on words in the pattern with some modifications, such as changing some grammatical meaning, part of speech, reordering words. Gathered from the Web, sentences are transformed into tests based on the patterns. There are two types of tests: multiple choice and error detection. All tests were evaluated by experts and 77-80% were regarded as "worthy".

Shei (2001) presents the concept *FollowYou!*, which transforms a raw text into language lessons, giving the student an opportunity to read his/her favourite articles with textbook-support. The learner's vocabulary knowledge is tested and recorded in the Profile Manager, which decides which words should be included in the next lesson. The Lesson Generator extracts definitions of the chosen words from the Dictionary, the collocations, their synonyms, and example sentences from the corpus. To test the effectiveness of the



YK juhlii sekavassa maailmantilanteessa

Oulun palveluala Opisto on into mukana Puhu selkokieltä-kampanjassa. Koulun ruokasali seinällä on ohjeita, ___ selkokieltä kannattaa puhua. Suomen kieli oppii v… …alla suomi.

välillä
mukaan
kuinka

– Selkokieli oli jo ennen kampanjaa tuttu koulu henkilökunnalle.

kampanja
- campaign
- crusade
- drive
- run
- cause
- effort
- movement
- political campaign
- whispering campaign

Figure 1: Story practice mode: exercises presented randomly from text.

lesson and to update the user's vocabulary model, some exercises need to be solved, e.g., gap-filling exercises. The main idea behind that project is that authentic materials—created by and for native speakers—are essential for the language learner.

## 3 Main principles and features

The main principle of our project is stimulating active language use in the process of learning from a text. By this we mean *active production* of required language forms while reading texts, rather than passive absorption of language examples or rules. We focus on learning the grammar as well as the vocabulary. Exercises provided by the system—including multiple-choice quizzes for indeclinable parts of speech, crosswords automatically generated from stories, can be regarded as grammar and vocabulary practice because the learner needs to produce words in context. Flashcards are available for vocabulary learning.

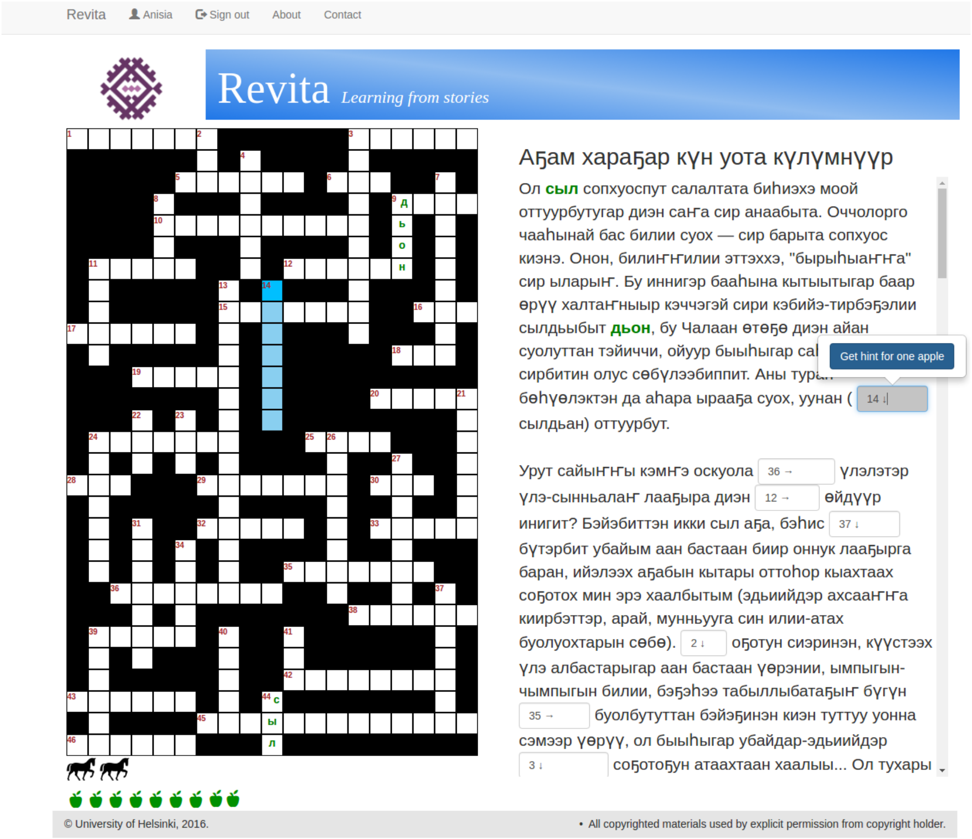
The platform has a small library of stories for each language. However, the main idea is that students will upload a variety of texts from web pages or plain text files to their personal library. Personal libraries can be shared between users. Studying language by reading stories, in which the students are interested implies personal involvement in learning process, it reduces boredom factor, and increases motivation to use the online platform. Moreover, texts uploaded from the Internet and mostly intended for native speakers will catalyze cultural enrichment and immersion into the specifics of language use and conventions.

One important system feature is that adding a new language is a simple procedure if a morphological analyzer is available for the language of interest. However, without language-specific adjustments and sets of rules, based on which the more complex exercises can be created, the kinds of available exercises will be limited and the range of grammatical concepts, which can be practiced, will also be restricted.

Exercises are created from any story automatically by analyzing words in the text and deciding on the best words to practice. The choice of words is based on the student's answers given so far, which the program remembers and assesses automatically. Tracking the students progress is one of the key features which we plan to develop during further research.

### 3.1 Essential exercise modes

There are two essential exercise modes provided by the system at present: the "practice" mode and the crossword mode. In the *practice mode*, see Figure 1, the learner chooses a story which s/he wants to practice and then receives pieces of this story in order. Each piece (called a "snippet")



Ађам харађар күн уота күлүмнүүр

Ол **сыл** сопхуоспут салалтата биһиэхэ моой оттуурбутугар диэн саҥа сир анаабыта. Оччолорго чааһынай бас билии суох — сир барыта сопхуос киэнэ. Онон, билиҥҥилии эттэххэ, "бырыһыаҥҥа" сир ыларыҥ. Бу иннигэр бааһына кытыытыгар баар өрүү халтаҥныыр кэччэгэй сири кэбийэ-тирбэђэлии сылдьыбыт **дьон**, бу Чалаан өтөђө диэн айан суолуттан тэйиччи, ойуур быыһыгар са сирбитин олус сөбүлээбиппит. Аны тура бөһүөлэктэн да аһара ыраађа суох, уунан ( 14 ↓ сылдьан) оттуурбут.

Урут сайыҥҥы кэмҥэ оскуола 36 → үлэлэтэр үлэ-сынньалаҥ лаађыра диэн 12 → өйдүүр инигит? Бэйэбиттэн икки сыл ађа, бэһис 37 ↓ бүтэрбит убайым аан бастаан биир оннук лаађырга баран, ийэлээх ађабын кытары оттоһор кыахтаах сођотох мин эрэ хаалбытым (эдьиийдэр ахсааҥҥа киирбэттэр, арай, мунньуура син илии-атах буолуохтарын сөбө). 2 ↓ ођотун сиэринэн, күүстээх үлэ албастарыгар аан бастаан үөрэнии, ымпыгын-чымпыгын билии, бэђэһээ табыллыбатађыҥ бүгүн 35 → буолбутуттан бэйэђинэн киэн туттуу уонна сэмээр үөрүү, ол быыһыгар убайдар-эдьиийдэр 3 ↓ сођотођун атаахтаан хаалыы... Ол тухары

© University of Helsinki, 2016.

• All copyrighted materials used by explicit permission from copyright holder.

Figure 2: Crossword generated from Sakha story

includes approximately 30-40 words, respecting sentence boundaries. Several words in the snippet will be chosen for quizzes, as the result of a randomized selection process. For each quiz word, the learner receives a gap in the text, and one of two types of quizzes: a multiple-choice quiz, where the learner must select one word from a list. Multiple-choice quiz is can be generated for non-inflected words, like prepositions, postpositions, adverbs, etc. The second type of quiz—cloze quiz—is used for inflected parts of speech: nouns, verbs, adjectives. The base form (lemma) is shown, and the learner needs to guess the correct grammatical surface form in the context. For example: "Topelius kertoo Maamme kirja eri maakunnista" (*"Topelius tells* book *Our Land about the different provinces."*). The word in the box is a lemma which is presented as a hint to the user. The task is to derive the surface form from this lemma in the given context. The correct inflected surface form in this example is "kirjassaan" (*"in his book"*).

After producing with all quiz words in the current snippet, the learner receives immediate feedback about his/her answers, and the next snippet for practice. The student receives points for correct answers, and points are removed if the user makes mistakes. It is important to stress that the correct form means the same as the form found in the story. This approach to assessment is convenient because we only rely on that the author chose to include in the story. However, it also has drawbacks because the user may insert a form which is allowed by the context but is not the same as the form used by the author in the story. This problem is one of the topics for further research.

Crosswords are generated from the story (or from a part of the story) automatically and consist of 40–50 words, see Figure 2. Users receive the story as an exercise, with some of the words removed, and a crossword based on the missing words. The task is to guess the words in their correct grammatical form. If the forms inserted by the user are correct, they will be added to the story and highlighted in green. Since this task can be difficult even for a native speaker, the user can request an additional hint for any missing word, which is its grammatical base form (lemma). The student receives points for solving the words.

During work on the current snippet, the student can request a translation of any word (more precisely, of its lemmas) in the snippet. The translation is shown in the box on the left, Figure 1. It is important to clarify the notion of *ambiguous* words in Revita. A word-form is considered as ambiguous if it has more than one different lemma. For instance, words with different, unrelated meanings can have homonymous forms but different base forms. For example, the Russian surface forms "жил" has two morphological bases: "жить" (live-INF, *"to live"*) and "жила" (sinew-NOM.SG, *"sinew"*). In the first case, "жил" is the past tense, masculine gender form of the verb (live-PST.MASC.SG, *"he lived"*), in the second case "жил" is the genitive plural form of the noun, (sinew-GEN.PL, *"sinew"*). If a word-form in the story is ambiguous, the system tries to provide translations of all base forms.

For Finnish, Swedish, and Russian, Revita uses the Glosbe multi-language dictionary[2] with a possibility to translate into a number of languages. FU-Lab dictionaries[3] are used to translate from Komi-Zyrian, Meadow Mari, and Udmurt into Russian. Revita uses *sakhatyla.ru* for translating from Sakha into Russian and English. The default destination language for translation will be the same as the language chosen by the user as the language of the interface (currently English, Finnish, Swedish or Russian) if dictionaries for these languages are available. For instance, for Komi-Zyrian, Udmurt, and Meadow Mari, translation is available only into Russian at the present stage. All words that the student has clicked on to get translations are automatically saved to the personal dictionary. Words the dictionary are used for practice as *flashcards*, with the lemma on one side of the card and its translations on the other side.

### 3.2 Generating exercises

Any uploaded text is first tokenised, the title is identified and the text is analysed by a morphological analyser. Revita uses the following tools:

- morphological analysers for Uralic languages, from GiellaTekno[4];
- the Crosslator Tagger (Klyshinsky et al., 2011) morphological analyzer for Russian;
- the HFST toolkit[5] for analyzing Swedish;
- *sakhatyla.ru*,[6] morphological analyser of online Sakha-Russian-Sakha translator system.

We extract base forms, parts of speech, and grammatical tags from the morphological analyses. Split into words and analysed, stories are saved into the database.

After morphological analysis, the system extracts from the text all words and combinations of words which can serve as candidates for practice. Every candidate is assigned to a particular snippet of the story and saved in the database. To be chosen as candidates, singleton words should have the same base form for all analyses returned by the analyser, otherwise, a word cannot be used for practice because the system cannot decide what base have to be offered as a hint. Combinations of words are chosen by the system based on language-specific rules; all words in a combination are considered to be disambiguated.

Choosing only unambiguous singleton words as candidates is a problem for the system because it limits the range of words and grammatical concepts which can be presented in exercises. For example, Udmurt forms in reflexive voice are homonymous to present tense forms, e.g., the verb "дасяны" (prepare-INF, *"to prepare"*) has a form "дасяськo" (prepare-PRES.3.SG, *"s/he prepares"*) with the meaning of singular present tense, and another verb "дасяськыны" (prerare-INF-REFL, *"to prepare oneself"*) has the homonymous form "дасяськo" (prepare-PRES-REFL.3.SG, *"s/he prepares her/himself"*), where the latter form has the meaning of reflexive voice. It means that the form "дасяськo" is ambiguous (has two different lemmas) and will never be chosen as a candidate by Revita. Consequently, the reflexive voice cannot currently be practiced for Udmurt for words with the same paradigm.

Combinations of words are chosen by Revita based on language-specific rules. For instance, the system contains rules for Russian, such as:

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

The rules make reference to the word's parts of speech and morphological tags. The first

[2] https://glosbe.com

[3] http://dict.fu-lab.ru

[4] http://giellatekno.uit.no

[5] https://kitwiki.csc.fi

[6] *Sakhatyla.ru* is created by Vasiliy Ivanov and has a web-interface available here[7] and a Telegram messenger application bot.

rule defines agreement between a noun-adjective pair. The second rule defines prepositional government—which cases are governed by the specific preposition, в ("in"). These rules drive the selection of sequences of words from the story, such as "красивой девушке", "в доме", which correspond to the specified rules as follows:

"красивой девушке"
  beautiful-Fem.Dat.Sg girl-Fem.Dat.Sg
  *"... [to] a beautiful girl (dative)"*

"в доме"
  in house-Loc.Sg
  *"In a/the house"*

Any possible ambiguity in the sequences matched by the rules is expected to be resolved[8] by virtue of the context. Sequences selected (randomly) by these rules will be offered as quizzes for practice as cloze-type exercises—the learner again receives as a hint only the lemmas of these words—or as multiple-choice quizzes. In case the sequence includes indeclinable words (such as a preposition, in the second rule, above) other prepositions with similar meaning will be used as distractors. Depending on the learner's results on other tasks, the system will offer exercises of various levels of complexity. For example, for sequences matching the above rules, we may produce:

- multiple-choice quiz for a preposition, all other surface forms given;
- one inflected surface form as cloze quiz (only the lemma given);
- one inflected word is as cloze quiz, multiple-choice quiz for a preposition;
- both noun and adjective surface forms as (coordinated) cloze quizzes, and multiple-choice for a preposition.

All of the learner's answers are stored in the database, both correct and incorrect. The entire history of the learner's answers is used for selecting exercises in subsequent snippets. Revita uses the history to compute weights for exercise candidates—non-ambiguous singleton words, and sequences of words that match rules. Examples which never always answered correctly by the learner receive a low probability (so they are not chosen frequently, to avoid boring the learner). Examples which were answered sometimes correctly and sometimes incorrectly receive high probability. Examples that were never answered correctly receive a lower weight again. Any time when the user starts practicing a new snippet, a probability of next candidates for practice is calculated. The system also controls the spread and proximity of the candidates within the snippet—they should not be too close to each other to provide sufficient context for each exercise. This randomness is applied when choosing from the set of all candidates—this allows each story to be practiced *multiple times*, with new exercises being chosen on each round. When the learner starts over, the system will select a new set of words for practice, which may partially overlap with the set of words chosen on the previous round.

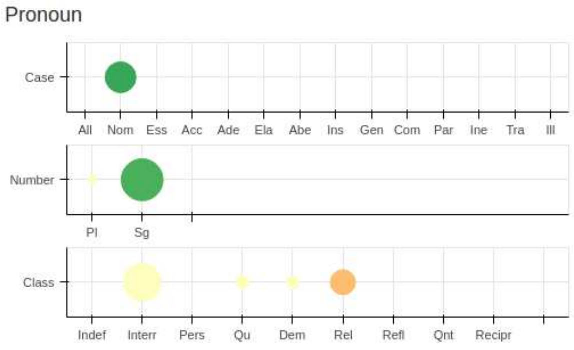


Figure 3: Progress visualisation for Finnish pronouns

At the current stage, the system provides an initial version of the learner's progress assessment. Revita checks all answers which the learner has provided during the exercises, and identifies which grammatical concepts were answered correctly what proportion of time; the concepts include grammatical categories, such as case, number, tense, etc. The learner (or teacher) can track progress via a visualisation page, which displays how the user performed on various concepts, see Figure 3. The more a grammatical concepts has been exercised the bigger its circle; the color ranges from green for mostly correct answers to red for mostly incorrect ones.

### 3.3 Code-switching disambiguation

Choosing words for exercises needs some care for certain languages, where a special kind of ambiguity arises. For example, texts in many of the F-U languages often include instances of *code-switching* into Russian. Code-switching is a normal and common phenomenon; however, only

[8] It is possible to construct (somewhat artificial) examples, where ambiguous words match these syntactic patterns and yet do not form the expected construction. If needed, this problem can be alleviated by various NLP techniques—by taking *wider* context into account.

words from the target language[9] should be chosen for practice. The problem arises when a Komi text has a surface form, *X*, which is a code-switch into Russian, and yet *X* happens to be **also** a valid word-form in Komi (with an unrelated meaning).[10] For example, *X* may be the surface form "пота". In Komi it is first-person singular indicative of the verb "потны" ("to crack"). The same word-form also happens to be the genitive of Russian "пот" ("sweat"). If we ignore the Russian, and *X* happens to be an instance of a code-switch (a Russian phrase inserted into the Komi text), then Revita will provide the Komi verb stem "потны" as a "hint."

In general, the clear danger is that Revita may incorrectly treat *X* as a Komi word-form, extract its Komi lemma and inappropriately offer the lemma as a "hint" to the learner in a cloze quiz—this would be terribly misleading, causing the system to lose credibility with the user.

To prevent this type of mistake, several methods may be applied. We present a simple solution, which works well for the present.[11] We apply morphological analysers for both Komi and Russian to *all* words in Komi text. If a word has only a Komi analysis, it becomes a candidate for exercises. If it has only a Russian analysis, it is definitely excluded as Russian. The last case is when the word has both analyses. We don't want to simply remove all such words from the list of candidates for exercises.[12] Thus, we apply this algorithm to identify and discard *"risky"* Russian words:

- for all words *w* with both Russian and Komi analyses;
- we look through the entire text and check whether *w* has *"friends,"* i.e., whether its base form is equal to the base form of some *other* surface form *y* in the story. We check this property, because we expect Komi words to repeat in the story. All words without friends are discarded as risky—they are potential Russian words mistaken as Komi. If *w* has Komi friends in the story, it is highly likely to be a true Komi word.
- If *w* has friends, we examine its *"neighbors."* The word is again discarded as risky if it has at least one direct neighbor with a Russian analysis, because we expect that Russian words are more likely to appear as part of entire phrases than as isolated words.

To evaluate the accuracy of the algorithm, we took a sample of 5% of all words having both a Russian and an Udmurt analysis and computed the accuracy of the prediction made by the algorithm:

$$accuracy = \frac{TP + TN}{all}$$

where *TP* are true positives—words marked as Udmurt by the algorithm, which an expert confirmed to be Udmurt. *TN* are true negatives—non-Udmurt words which the algorithm marked as Russian. We manually checked the sample of words with Russian and Udmurt analyses in our corpus of stories. The obtained accuracy was 0.77. We should note that Crosslator Tagger sometimes returns a Russian analysis for non-Russian words, which increases the number of false positives (words which are not really risky), and brings down the accuracy measure.[13]

Because we expect the learner to produce the grammatical form which is equal to the form found in the story, we assume that there is only one correct answer in a particular context. However, we can have lexical and grammatical synonyms which suit the same context, as well as *optional* grammatical meanings which may or may not be expressed in this context, which may make it difficult for the user to guess the correct grammatical form only from the lemma. The system should not choose such cases for practice or should be more intelligent and tolerate optional or grammatically equivalent markers. For instance, in Komi-Zyrian the same grammatical meaning can have different forms, e.g., verb "лоны" (*"to be"*) in the indicative mood, first past tense, third person singular has two valid forms with the same meaning in the same context — "лои" and "лоис". Thus, the learner cannot decide which form is expected by the system. Solving that problem is non-trivial because it requires sufficient amounts of data to build a reliable language model. We plan to start with

[9] In this section we will refer Komi as a "representative," to avoid writing repetitively "*a F-U language that uses the Cyrillic alphabet and therefore may contain word-forms confusable with Russian*."

[10] Note, this does not apply to *borrowings*, where Russian words are borrowed into Komi, and inflected according to Komi morphological rules.

[11] More robust and ultimately better solutions will involve building statistical language models, planned for future work.

[12] In Udmurt, e.g., they represent 19% of all words in our corpus.

[13] We have tested only Udmurt, we will test with other languages which exhibit code-switching into Russian.

Finnish and Russian because for other languages data is more difficult to obtain.

## 4 Conclusions and future plans

The Revita system is under development, therefore there are many outstanding problems to be solved and improvements to be added. Continuing the above discussion, presenting an appropriate hint to the learner is crucial, because misleading hints (lemmas for cloze quizzes and distractors for multiple-choice quizzes) will cause the learner frustration and will discourage the continued use of the system. Further, we must solve many language-specific problems. While for some languages, like Russian or Finnish, it may be done by building language models, for languages like Erzya, Komi-Zyrian, or Sakha we may have to develop rule-based solutions, due to a lack of corpora. Also, many difficulties are caused by erroneous analyses. We discussed the process of generating exercises at the current stage and related problems. Further kinds of exercises can be developed for different languages depending on the available language resources.

The system was tested by several users, and we plan to collect more formal results about its efficacy.

Revita offers several types of exercises generated from any story. The systems assesses the answers given by user by comparing them with forms found in the story and it cannot accept other answers which are allowed in the context.

Users can translate any word in the story and to save them as flashcards. Based on the flashcards, Revita provides vocabulary exercises. Vocabulary learning in general and vocabulary learning with help of computers was studied, e.g., by (Nation, 2013), (Ahmed, 1989), (Laufer and Hill, 2000), (Prince, 1996). Learning new words in context is more preferable than learning words in isolation—see (Groot, 2000) and (Krashen, 1989)—to better understand their semantic and syntactic features. This is consistent with one the main principles of the system, namely, learning language while reading. The learner does not only infer the meaning of a new word from the context, but also can link it with a translation into the learner's native language. Efficiency of such linking is questioned, despite the efficiency in terms of quantity, see (Prince, 1996). Nevertheless, we assume this linking to be beneficial provided that there are other approaches to learning offered in parallel. This may involve establishing links between a new word and other words in the language, e.g., through exercises with synonyms, where the learner should decide which word among a list of synonyms is the most appropriate in the context, and to generate the correct grammatical form of the chosen word. This type of exercise can also include practicing of multi-word expressions.

Further aspects which we plan to develop are:

- refining the scoring system which should not "only lead to a learner's pursuit of meaningless 'points' with little or no regard for learning" (Beatty, 2013) but works to stimulate the user to learn more;
- adding the possibility for collaboration to the system, since some of the pedagogical objectives can be achieved better through group activity—solving problems in a group, discussing them with experts/teachers also registered in the system.
- assessment of uploaded stories by their difficulty for the learner, and their quality as learning material. This is important because the learner decides which stories to practice, and the system should help guide learners in some may.
- progress detection which is important for developing new exercises and their assessment.

Progress detection and assessment involves comparing previous responses of the user and identifying the development of his/her knowledge, targeting weak areas, and generating exercises for the next stage, depending on all this information, and returning intelligent and useful feedback to the learner.[14] Development of this functionality is one of the main future steps in the Revita system.

## Acknowledgments

This work was supported by the FinUgRevita Project of the Academy of Finland. We thank Larisa Ponomareva for help with the Permic languages, Sardana Ivanova for help with Sakha, Max Koppatz and Kim Salmi for their contributions to the software development.

[14] This is a challenge, since we wish to avoid assuming that the learner is familiar with any linguistic or grammatical concepts; the system should serve non-specialists equally well.

## References

Medani Osman Ahmed. 1989. Vocabulary learning strategies. *Beyond words*, pages 3–14.

Ken Beatty. 2013. *Teaching & researching: Computer-assisted language learning*. Routledge.

J. Brown, G. Firshkoff, and M. Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of HLT/EMNLP-2005*.

Carol Chapelle and Joan Jamieson. 1983. Language lessons on the PLATO IV system. *System*, 11(1):13–20.

C. Chen, H. Liou, and J. Chang. 2006. Fast—an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL, Interactive Presentation Sessions*.

Larrie E Gale. 1989. Macario, Montevidisco, and interactive Dígame: Developing interactive video for language instruction. *Modern technology in foreign language education: Applications and projects*, pages 235–247.

Peter JM Groot. 2000. Computer assisted second language vocabulary acquisition. *Language Learning & Technology*, 4(1):60–81.

Robert Hart. 1981. Language study and the PLATO system. *Studies in Language Learning*, 3(1):1–24.

Trude Heift. 2001. Intelligent language tutoring systems for grammar practice. *Zeitschrift für Interkulturellen Fremdsprachenunterricht*, 6(2).

ES Klyshinsky, NA Kochetkova, MI Litvinov, and V Yu Maximov. 2011. Method of POS-disambiguation using information about words co-occurrence (for Russian). *Proc. of GSCL*, pages 191–195.

Stephen Krashen. 1989. We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The modern language journal*, 73(4):440–464.

Batia Laufer and Monica Hill. 2000. What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? *Language Learning & Technology*, 3(2):58–76.

John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Proceedings of INTERSPEECH*, Antwerp, Belgium.

Denise E Murray. 2014. *Knowledge machines: Language and information in a technological society*. Routledge.

Noriko Nagata. 2002. Banzai: An application of natural language processing to web-based language learning. *CALICO journal*, pages 583–599.

Ian Stephen Paul Nation. 2013. *Teaching & learning vocabulary*. Boston: Heinle Cengage Learning.

Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains at the 9th Internationnal Conference on Intelligent Tutoring Systems*.

Peter Prince. 1996. Second language vocabulary learning: The role of context versus translations as a function of proficiency. *The modern language journal*, 80(4):478–493.

Chi-Chiang Shei. 2001. FollowYou!: An automatic language lesson generation system. *Computer Assisted Language Learning*, 14(2).

Simon Smith, P V S Avinesh, and Adam Kilgarriff. 2010. Gap-fill tests for language learners: Corpus-driven item generation. In *Proceedings of ICON-2010: 8th International Conference on Natural Language Processing*.

W. L. Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, 30.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 143–148, Baltimore, Maryland, USA.