Digital cultural heritage and revitalization of endangered Finno-Ugric languages

Anisia Katinskaia and Roman Yangarber

University of Helsinki, Finland
first.last@cs.helsinki.fi

Abstract. The preservation of linguistic diversity has long been recognized as a crucial, integral part of supporting our cultural heritage. Yet many "minority" languages—those that lack official state status—are in decline, many severely endangered. We present a prototype system aimed at "heritage" speakers of endangered Finno-Ugric languages. Heritage speakers are people who have heard the language used by the older generations while they were growing up, and who possess a considerable passive competency—well beyond the "beginner" level,—but are lacking in active fluency.

Our system is based on natural language processing and artificial intelligence. It assists the learners by allowing them to learn from arbitrary texts of their choice, and by creating exercises that engage them in active production of language—rather than in passive memorization of material. Continuous automatic assessment helps guide the learner toward improved fluency. We believe that providing such AI-based tools will help bring these languages to the forefront of the modern digital age, raise prestige, and encourage the younger generations to become involved in reversal of language decline.

1 Introduction

The rapidly developing computational technologies are expanding into every domain related to languages. Computers are used as tools to support language learning and maintenance. During the last decade, many online, widely accessible language learning tools have emerged. However, most of them do not cover or support minority languages.

In this paper, we introduce Revita—a freely available online platform, designed to support learning/tutoring for endangered languages, *beyond the beginner level*. Revita currently works for several endangered or minority languages: Udmurt, Meadow Mari, Erzya, Komi-Zyrian, Komi-Permiak, North Saami, and Sakha.[1] All of these are Finno-Ugric languages, except for the last, which is a Turkic language. Most of the languages are represented inside the Russian Federation (RF), each with a moderate-to-small number of speakers. The system also works for several "major" languages—currently, for Finnish,

[1] The system can be viewed at `revita.cs.helsinki.fi`

Swedish, German, Russian, and Kazakh. This functionality—support for major languages—evolved in Revita in part for practical reasons. For instance, Finnish is closely related and structurally similar to other Finno-Ugric languages. The Russian language exerts a powerful influence on most of the above-mentioned languages, since they are represented primarily inside the RF, where Russian is the only state official language. Most communication (including written) in these languages exhibits common, spontaneous code-switching into Russian.

Revita can automatically generate a large variety of exercises, created from arbitrary, real texts that are chosen and uploaded by the learners themselves. Alternatively, the texts can be uploaded by teachers, or shared by learners with one another. The key research challenge underlying the system is the aim to adapt the level of exercises to each individual user, depending on her level of competence. The system tries to estimate the level of competence based on the learner's answers to exercises. In this way, Revita lies at the intersection of two well-established research areas: Intelligent Tutoring Systems (ITS) and Computer-Assisted Language Learning (CALL).

This paper is structured as follows: Section 2 presents prior work in computational tools for language learning. Section 3 describes the system in greater detail, and positions Revita in the field of Educational Data Mining (EDM), and Section 4 presents conclusions and pointers for future work.

2 Prior work

The idea of using computers to support language learning has emerged over 50 years ago. The research field of CALL—Computer assisted language learning—encompasses a broad variety of technologies, which serve to support language learning and teaching. CALL is briefly defined as "the search for and study of applications of the computer in language teaching and learning," [7]. One of the first CALL systems, PLATO, was created in the early 1970s [4] and today one can find many systems, most of them commercial, and most aimed toward beginners and focusing on a limited set of (commercially popular) languages.

The idea of using computer assisted language learning for endangered languages is not widespread, and in the field of revitalization, CALL is a relatively new concept. A number of studies investigate how technology can influence the revitalization process, [1,11,13,12,15].

With the increase of accessibility of new technologies, the methodology of language revitalization is also evolving. For example, tools for recording and sharing language data collected from native speakers in authentic contexts are accessible, and the tools facilitate the tasks to a much greater extent than was possible previously. However, we still encounter skeptical views relative to the use of computational technologies in teaching languages. [11] recognizes the effectiveness of computer as a tool for collecting data, but also underlines that while the computer provides new opportunities for teaching, the new technology creates new pedagogical demands for its effective implementation. The

main concern is the threat of an absence of authentic communicative environment in settings where language learning is aided by computer.

Nevertheless, with the rapid expansion of the Internet, computational technology is receiving wide application for revitalization purposes. The Hawaiian language is a good example of how technology can assist in supporting languages. In preservation and dissemination of language materials and developing multiple models of communication, technology played a most significant role for Hawaiian, [14].

For example, the Ulukau website [2] provides access to valuable resources for teaching and revitalization. This is important, considering the shortage of texts available for Hawaiian. The Leoki bulletin system, [3,16], provides communication by email, online chats and conferences, announcements about the Hawaiian language, online order forms for the purchase of Hawaiian language books, dictionary databases, where users can suggest new words, issues of newspapers, posted stories and songs, information about educational support, and so on. Leoki has provided an opportunity for communication among speakers separated by distance, which is crucial, since for some students e-communication is the only chance to use the language outside of classroom.

Some application and platforms are already in use for supporting endangered languages. Memrise[3] is a learning platform for courses created by users, and it includes several courses in Irish and Welsh. Chickasaw Language Basic[4] is an application for Apple mobile devices for learning Chickasaw, a Native American language of the Muskogean language family. It offers videos, songs, words and phrases in Chickasaw. Aikuma[5] is an application for collecting data from endangered language speakers. The users can easily make records, enhance them with meta-data, offer translations phrase-by-phrase and to share it with other users of the app. Tusaalanga[6] is an iOS application for learning of five Nunavut dialects (spoken in Northern Canada), which has dialogues with audio, grammatical lessons and glossaries with audio. The "Ma! Iwaidja" dictionary[7] is an application for learning Iwaidja, an Australian language. Any user can insert new words, phrases, and their translations. The Skidegate Haida Language application[8] for Haida, which is spoken by the Haida people in the Haida Gwaii Archipelago, off the coast of Canada, and on Prince of Wales Island in Alaska. This application has a bilingual dictionary and collection of phrases. Words and phrases are illustrated by pictures and audio. Users have an option to edit the content and replace it with their own images and audio recordings. "Learn Manx"[9] is an application for Manx, a Goidelic Celtic lan-

[2] http://ulukau.org
[3] https://www.memrise.com/
[4] https://itunes.apple.com/us/app/chickasaw-language-basic/id448797486?mt=8
[5] http://www.aikuma.org/aikuma-app.html
[6] http://www.tusaalanga.ca/ios/about
[7] https://play.google.com/store/apps/details?id=com.pollen.maiwaidjadictionary
[8] http://www.firstvoices.com/en/Hlgaagilda-Xaayda-Kil
[9] https://play.google.com/store/apps/details?id=com.anspear.language.manx

guage, which has become extinct as a first language, but has around 2000 speakers. The application includes words and basic phrases, bilingual dictionaries, a flashcard learning system; it allows the recording of responses to questions to compare with model responses, and exercises on grammar and comprehension. Most of these applications have some language materials; they often offer the possibility for users to add more data from informants and include dictionaries and limited set of phrases to learn.

Several popular *commercial* language learning systems cover languages considered to be endangered. For example, Duolingo offers courses for learners of Irish and Guaraní, an indigenous language of South America. Rosetta Stone, a commercial language-learning software provider, has established the Endangered Language Program,[10] whose goal is to revitalize several endangered languages. The program claims to provide support for Chicksaw, Mohawk, Chitimacha, Inuktitut, Inupiat, and Navajo. (However, we do not find these languages in the list of languages available for practicing by a registered user on the platform.)

We do not draw a clear distinction between using CALL for language maintenance in general vs. for teaching/learning endangered languages and revitalization of "heritage" languages. The latter are languages spoken by people whose ancestral language can be considered indigenous, may lack official status in the area where it is spoken, and may also be endangered, [10]. In any case, whether the heritage language is endangered or not, the learner uses it regain or retain access to the ancestral culture linked to this language.

A number of studies examine ways in which CALL can be helpful for learning heritage languages. For example, in [8], the authors examine how computer-mediated communication (CMC) has helped Russian heritage speakers in the USA in the acquisition of academic-level literacy. CMC includes all forms of communications mediated by computer: email, forums, chat-rooms, messages, etc. All of these forms of communication with instructors and other learners can help improve writing and reading in the heritage language, in a range of registers. As the authors stress, this can help to observe the target language in use, access relevant resources about the language, and to anchor the oral language—with which learners are more familiar—in the written form. The use of CMC was shown to have a positive effect on vocabulary acquisition, spelling skills, composing messages, "spoken" writing, grammatical competence, and attention to punctuation. More communication in the target language also entails a growth of interest in the heritage culture, and in exploring the cultural identity of the learners.

Another example of a CALL system for learning a heritage language is described in [6]. The system provides exercises to learners of Runyakitara, a Bantu language. The language is not endangered, as it is spoken by over 6 million people in Western Uganda. The system is aimed at native speakers of Runyakitara, but with limited competence in this language. Usually these are children of Runyakitara migrants who have a very rudimentary knowledge of the

[10] https://www.rosettastone.com/endangered

language. The aim of the project is to introduce such learners to their native language, develop literacy skills and increase the learners' respect toward and pride in their culture. The main focus is noun morphology, which is difficult to learn. The system for practice includes morphological exercises and testing of the learners' knowledge of morphology and vocabulary; it also provides scores which can help teacher to evaluate the learners' progress. All nouns in the system were extracted from a Runyankore-Rukiga dictionary, Kashoboorozi, and parsed by a finite-state morphological analyzer.[11] The system offers exercises, such as plural forms of nouns. The user should type in an answer and can receive feedback about whether the answer was correct. The learner can also get supplementary material for grammatical explanations. The system saves information about the user, including the dates of practice sessions, the material covered and the scores. The teacher can obtain information about scores of all learners by lessons. Results of pre-testing and post-testing (after completing all exercises) showed a significant increase in grammar scores. Experiment with the system proved its effectiveness for the task of learning the heritage language by its native speakers, a majority of whom reported that they would like to continue using it.

3 System description

3.1 Features of the Revita system

In this section we describe the main features of Revita language learning system, [5]. The system is developed based on the idea of providing users the possibility to learn languages actively, rather than passively absorbing learning materials. This active model of learning has broad implications and manifests itself in the following:

- Users seek out learning materials, which are of interest to them. In this way the learner collects a database of materials and exercises, which can be useful for pedagogical purposes for endangered languages.
- The user develops her active language skills by *producing* language forms in the context of the story; in most cases, exercises involve unrestricted language forms. (In some settings—such as mobile use, where typing may be less convenient, the system may offer multiple-choice exercises.)
- The feedback provided should offer more insight than simply "correct" vs. "incorrect". Further, rather than revealing the correct answer immediately after the first failed attempt, a more clever approach should push the student further to seek out the correct answer by herself.
- Advanced modes of using Revita system (which are currently under development) expand on the idea of providing the possibility for advanced users—users who have some competency also in linguistic terminology—to

[11] This approach is relevant in our context, since Revita also makes use of finite-state morphological analyzers as low-level supporting components.

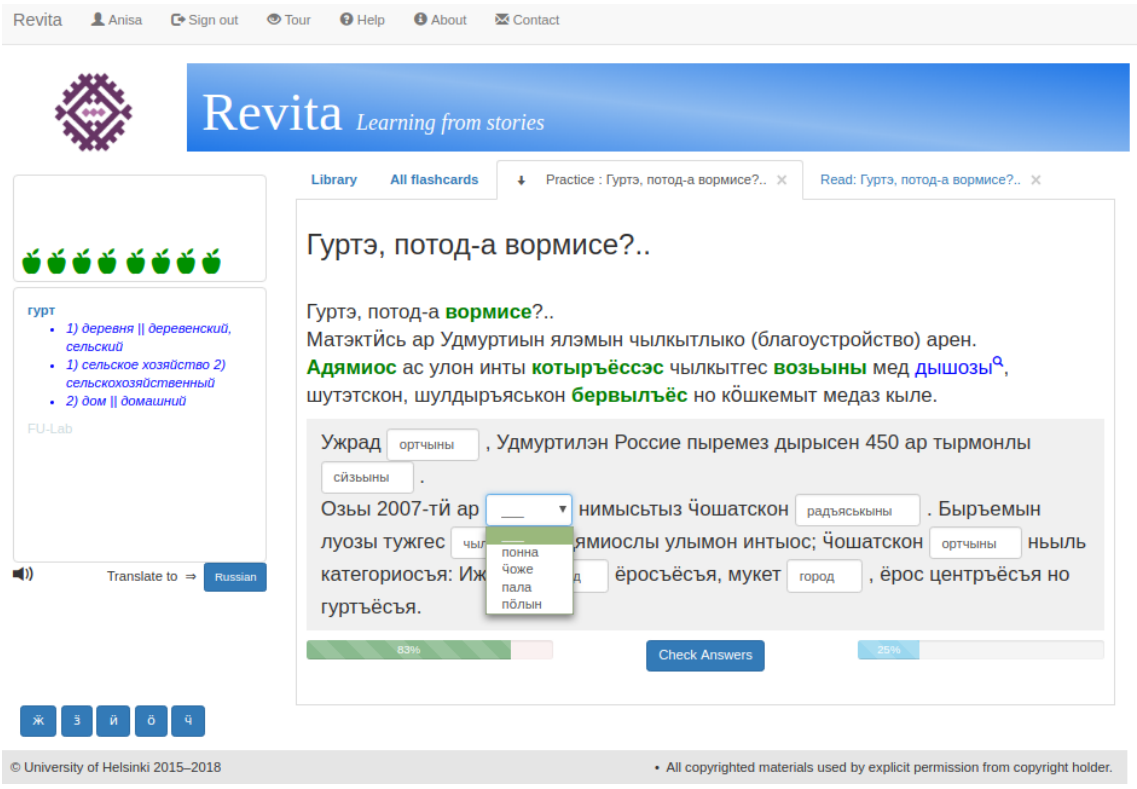

Fig. 1. Story practice mode for the Udmurt language.

filter exercises by linguistic *concepts*, to monitor and direct their own learning progress more actively and on a finer level.

The system has a small "public" library of stories for every language, which are available for all users, including non-registered users. It can be problematic for users to find suitable learning material when beginning to work on learning a language; the system offers some texts to begin practicing with stories. Revita also offers links to sources of other authentic texts, including newspapers, popular journals, etc., as well as and information about the language. We plan to extend the public library for all languages with materials and links supplied by experts in the respective languages.

A central idea behind the platform is that users will *actively add their own* learning materials, to their private, personal libraries. Texts can be uploaded from a personal computer (.txt or .doc files), or by copying-pasting from other sources, or by loading from a Website—the user provides a URL of a Web page containing text material she wishes to use for practice. Users can also upload stacks of flashcards containing words to practice, with their translations or definitions. In future we plan to extend these types of learning materials with audio

data. For endangered languages this option can be very important because then the system will also serve as a tool for preserving language data, as many CALL systems do. Uploaded materials can be shared with other users.

We consider the possibility to practice using material chosen and uploaded by the user as a key feature of the Revita system, because such material is particularly suited to keep the user interested and motivated. Another aspect which makes the practicing process more engaging is that it employs real-word texts, rather than artificial texts specially constructed for exercises. Such "natural" materials can better help to get immersed in contemporary culture and the life of the speakers of the target language. There are several exercise modes available to users. Without registration users can read stories in the *reading mode* and do exercises in the *practice mode*, but their results are be saved and cannot be used for adapting future exercise sessions to their competence level.

The reading mode is simplest; it allows the learner to become acquainted with the story, and to ask for translations of unfamiliar words in context. The practice mode is currently the main type of exercise mode in Revita, see Figure 1[12] The learner chooses a story to practice and receives it piece by piece, with some of the words *obscured* for *cloze* exercises. When each story is uploaded, it is analyzed by several natural language processing (NLP) modules (including morphological analyzers), and all possible candidates for exercises are chosen and saved. The choice of exercises in a given session is random, but it depends on the previous answers given by the user: exercises which were easy or too difficult to answer are chosen for the new exercises with lower probability, in order not avoid boring the student (with questions that are too easy) or discouraging the student (with questions that are too difficult) too frequently.

The user can receive two types of exercises in the practice mode: multiple choice—more suitable for non-inflected words—or cloze quizzes for inflected parts of speech. In the latter case, the user receives the base form of a hidden word, and needs to insert a correct grammatical form of the word appropriate for its context. The correct answer is the answer appearing in the story. The possibility to insert different forms acceptable in the same context is under development. The same story can be practiced more than once, as the generated exercises will be differ from the previous ones, because they influenced by the history of the previous answers.

After answering questions, the user receives immediate *feedback* and the next piece of story with new exercises. We plan to develop the feedback functional-

[12] The user has opened this story in the Practice mode (in another tab, the a story is open in the Reading mode). The highlighted snippet contains the current exercises. Exercises are of two kinds: white boxes contain base forms of words, where the user must type in the inflected form correct for the context; drop-down menus are multiple-choice questions. The snippet above was answered previously—correct answers are in green, incorrect in blue, with a magnifying glass to allow inspection of the "mistake." Each correct answer gives one point (an apple) in the score box. Any word can be clicked to obtain a translation—in the left panel, into a choice of languages. In lower-left are buttons for entering symbols not found on common keyboards. Progress bars indicate proportion of correct answers (left) and story covered (right).

ity further in such a way that it is not only providing a correct answer, but a series of hints which can help guide the user toward the solution after several attempts, and adapt to the particular user.

For another type of exercise, the system can generate *crosswords* from the stories. Entries in the crossword are chosen following the same principles as for the practice mode, considering the history of previous answers. A crossword is built from the hidden words which should be inserted back into the story in the correct grammatical form. The user receives the *translations* of the required words as hints; in this way, the user actively expands both the vocabulary and grammatical competency.

A *competition* feature is available for some exercise modes, where while practicing with a story, the user is trying to beat an opponent who is doing the same exercises based on the same story. This injects *timing* and speed into the practice sessions, where the learner has additional constraints due to the opponent's advance through the session. Currently, the opponent is a *bot* that attempts to model as closely as possible the user's own recent answering and timing patterns. Thus, the idea of this mode is for the user to compete exactly with *herself*—trying to improve on the correctness and speed of her answers.

During the practice sessions, the user can request *translations* for (most) unfamiliar words in the text—all such requests are automatically added into the user's set of flashcards, as we assume that these words are good for the user to review later. Flashcards can be used for vocabulary practice sessions apart from reading the stories. (Flashcards are also a common feature in some language-learning platforms.)

3.2 Small vs. big languages

One of the main advantages of Revita system is that it is relatively easy to add a new language, if a morphological analyzer and any other required NLP modules are available for that language. With such modules and some assistance from language experts for making required testing and adjustments, Revita can be used to generate exercises from any uploaded story in then new language. Some of the more advanced features depend on the availability of large quantities of language data for developing more robust language models. Finding sufficient quantities of data can be problematic for some of the smaller languages, thus limiting the capacity of the learning platform.

Thus, to some extent Revita provides different functionality for different languages depending on available language data.

4 Conclusions and current work

We have presented Revita, an on-line platform that allows us to explore several large-scale, important challenges recognized in digital humanities today.

In the area of cultural heritage, it helps us address the global problem of language endangerment—by bringing state-of-the-art AI tools used for learn-

ing "larger" languages to benefit endangered minority languages. Not coincidentally, by embodying an on-line, technological solution, it in part addresses the important sub-problem of raising prestige of the minority languages, by injecting these languages into the center of modernized discourse. It is well understood that prestige is a crucial social factor in language decline and endangerment.

We believe that our approach provides exciting new and rich sources of data for studying and modeling the process of language learning, with the aim of enhancing the learning experience. It opens opportunities for new research in educational data mining. We are currently investigating application of state-of-the-art methodologies in the context of Revita, including Bayesian knowledge tracing, [9] and knowledge space theory, [2]. In other scientific areas, such as CALL and ITS, it allows us to bring the latest advances in AI to bear on the modern understanding of pedagogical methodology.

Acknowledgments

This research was supported in part by the FinUgRevita Project, funded by the Academy of Finland, Grant No. 267097.

References

1. Buszard-Welcher, L.: Can the web help save my language. The green book of language revitalization in practice pp. 331–45 (2001)
2. Doignon, J.P., Falmagne, J.C.: Knowledge spaces. Springer Science & Business Media (2012)
3. Hale, C.: How do you say computer in Hawaiian? (1995)
4. Hart, R.: Language study and the PLATO system. Studies in Language Learning 3(1), 1–24 (1981)
5. Katinskaia, A., Nouri, J., Yangarber, R.: Revita: a system for language learning and supporting endangered languages. In: 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. Linköping University Electronic Press (2017)
6. Katushemererwe, F., Nerbonne, J.: Computer-assisted language learning (CALL) in support of (re)-learning native languages: the case of Runyakitara. Computer Assisted Language Learning 28(2), 112–129 (2015)
7. Levy, M.: Computer-assisted language learning: Context and conceptualization. Oxford University Press (1997)
8. Meskill, C., Anthony, N.: Computer mediated communication: tools for instructing Russian heritage language learners. Heritage Language Journal 6(1), 1–22 (2008)
9. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J., Sohl-Dickstein, J.: Deep knowledge tracing. In: NIPS: Advances in Neural Information Processing Systems. pp. 505–513 (2015)
10. Revithiadou, A., Kourtis-Kazoullis, V., Soukalopoulou, M., Konstantoudakis, K., Zarras, C.: Developing CALL for heritage languages: The 7 Keys of the Dragon. The EuroCALL Review 23(2), 38–57 (2015)

11. Villa, D.J.: Integrating technology into minority language preservation and teaching efforts: An inside job (2002)
12. Ward, M.: The additional uses of CALL in the endangered language context. ReCALL 16(2), 345–359 (2004)
13. Ward, M., Genabith, J.: CALL for endangered languages: Challenges and rewards. Computer Assisted language learning 16(2-3), 233–258 (2003)
14. Warschauer, M.: Technology and indigenous language revitalization: Analyzing the experience of Hawai'i. Canadian Modern Language Review 55(1), 139–159 (1998)
15. Warschauer, M.: Technology and social inclusion: Rethinking the digital divide. MIT press (2004)
16. Warschauer, M., Donaghy, K., Kuamoÿo, H.: Leoki: A powerful voice of Hawaiian language revitalization. Computer Assisted Language Learning 10(4), 349–361 (1997)