Support for endangered and low-resource languages via e-Learning, translation and crowd-sourcing

ROMAN YANGARBER

This article appears in:

Endangered languages and the land: Mapping landscapes of multilingualism

Proceedings of the 22nd Annual Conference of the Foundation for Endangered Languages (FEL XXII / 2018)
Vigdís World Language Centre, Reykjavík, 23–25 August 2018

Editors: Sebastian Drude, Nicholas Ostler, Marielle Moser


Cite this article:

First published: December 2018

Link to this article: http://www.elpublishing.org/PID/4014

This article is published under a Creative Commons License CC-BY-NC (Attribution-NonCommercial). The licence permits users to use, reproduce, disseminate or display the article provided that the author is attributed as the original creator and that the reuse is restricted to non-commercial purposes i.e. research or educational use. See http://creativecommons.org/licenses/by-nc/4.0/

Foundation for Endangered Languages: http://www.ogmios.org
EL Publishing: http://www.elpublishing.org
Support for endangered and low-resource languages via e-Learning, translation and crowd-sourcing

Roman Yangarber
University of Helsinki, Finland
Department of Computer Science
{yangarber.roman@cs.helsinki.fi}

Abstract
The REVITA project aims to create an automated Web-based environment for language learning, supported by state-of-the-art, scalable methods. The system is currently deployed for several low-resource languages from the Finno-Ugric and Turkic families. Our approach rests on four main ideas. 1: Leveraging existing NLP tools and resources. 2: Providing personalized learning, intended for learners at the intermediate-to-advanced levels. The learner can practice, receive feedback, and make progress on one’s own, outside the classroom. 3: Making the learning process maximally interesting, flexible and stimulating, to ensure longer practice time. 4: Engaging the learner in active production of language—following modern principles of didactics—rather than in passive memorization of linguistic material. In this paper we focus especially on crowd-sourcing and translation models for supporting e-Learning.

Keywords: e-learning, low-resource languages, translation, sequence-to-sequence models, crowd-sourcing

Introduction

The REVITA project aims to develop a system for language learning which is freely available and aimed at students beyond the beginner level.

Many resources exist on the Web for beginners, some with millions of users. However, once the learner passes the beginner’s level and reaches low-intermediate to advanced (LIA) level (above CEFR level A1/A2), available resources become drastically limited. As our research has shown (Katinskaia & Yangarber, 2018; Katinskaia et al., 2018), available systems do not provide substantial support for LIA learners in multiple languages.

The project, funded by the Academy of Finland, initially focused on languages from the Finno-Ugric language family. However, the system was designed with scalability and general applicability in mind. At present, it works with several Finno-Ugric and Turkic languages (Sakha, Kazakh), as well as several “major” Indo-European languages; German, Russian, Swedish. We plan to incorporate additional languages in the future.

The system crucially builds on pre-existing lower-level language technology resources. For a new language to be realized in the e-Learning platform, certain resources must exist—dictionaries, morphological analyzers, etc.—as detailed in Section 4. We leverage resources collected via documentation efforts, and natural language processing (NLP) research, to effect positive change in language learning and in active language use in speaker communities.

This paper outlines our approach to language learning building on the latest advances in artificial intelligence (AI) and NLP. Section 2, discusses the background behind computer-assisted language learning (CALL). Section 3, presents the main technical ideas and features of our system. Section 4, discusses leveraging existing NLP tools. Section 4.3, covers the benefits of engaging communities of learner. In Section 5, we discuss the current status and plans for further development.

Prior work

In computer-assisted language learning, computational technology is used for presenting learning material, practicing and assessment. Levy (1997) defines CALL as “the search for and study of applications of the computer in language teaching and learning.” The computer is intended to be a support tool, rather than a replacement for a teacher. CALL originated in the 1960’s, e.g., with PLATO (Hart, 1981) among the earlier systems. From the pedagogical perspective, its history can be divided into three phases (Warschauer, 2013):

- Behavioristic CALL (1960–1970s)
- Communicative CALL (1980–1990s)
- Integrative CALL (1990–present)

In the first phase, the computer was used for “drilling” learning material. It prompts the student with questions, and the student gives answers by filling in gaps (“cloze” quizzes), choosing an answer from a set (multiple-choice), by pronouncing the answer (“listen-and-repeat” exercises), etc. This approach is used today, and most CALL systems (free and commercial) exploit it in some form.

In the communicative phase, focus shifts toward using language, and learning grammar and vocabulary implicitly. Students are given tasks in the form of games; for example, text reconstruction exercises designed in the form of games where the student must help a character by inserting phrases describing the way home. However, creating a tool which can handle a variety of student responses and provide meaningful feedback is challenging.

The next phase integrates explicit (“drilling”) and implicit (e.g., tasks based on communicative situations) learning. It features multimedia systems, which present authentic discourse, allow the student to learn at her own pace and to control the path of interaction. A virtual game, e.g., may simulate various communicative situations, and include language exercises, and even crowd-sourcing tasks. Virtual reality (VR) also falls into this phase. One VR-based system, ImmerseMe,1 claims to prepare users for real-word

---

1 https://immerseme.co
communicative situations by exploration of over 500 scenarios in 9 languages.

CALL for less-resourced languages is an area of growing research, e.g., (Sharoff et al., 2014; Al Emran & Shaalan, 2014; Marítxalar et al., 2011). At present, most of the systems that we have surveyed (Duolingo, LinguaLeo, Babbel, Busuu, FluentU, etc.) support learning at the elementary/beginner level, and the exercises that the learners receive are “canned”—drawn from a pre-defined, fixed set, even if the set may be extensive. Exercises are created by human experts in advance of the study session, and are based on fixed text material, pre-determined by the designers of the exercises. They rarely cover context beyond a single sentence. Thus, these systems use a “one-size-fits-all” approach, offering a fixed set of exercises; even if the system does attempt to adapt to the learning process of the particular user, it does so only at the elementary level, and does not personalize future exercises to the user depending of her prior errors. They offer little of no support for LIA learners.

Main technical ideas

The main technical ideas behind our approach are as follows. First, REVITA complements—but does not replace—the human teacher. It provides a means for complementary practice outside the classroom, beyond in-class learning. Second, we aim to engage the learner, and to keep her maximally interested. Many textbooks, (including on-line ones), and learning platforms offer excellent materials, carefully pre-selected by competent teachers. Unfortunately—no matter how carefully constructed the materials—they offer no way to assure that the content is interesting to all learners, or, indeed, to any learner. The one-size-fits-all approach does not work in practice.

How does one assure that the learner is maximally interested? The solution we propose is to allow the learner complete freedom in choosing the learning materials. If I’m interested in science, or football, or horses, etc., I’m more likely to practice longer if I can use texts about my favorite subject, than if I’m “forced” by the platform to use canned material—static content, pre-selected by expert teachers. What is the source of the learning material? It should be the entire Internet, or any text in private possession of the learner. Content on the Internet is growing continually, which is fortunately true for many endangered languages as well.

The problem, then, becomes turning these arbitrary materials—chosen by the learner—into learning material. This is achieved by using NLP and AI techniques, as explained below.

Finally, we focus on active production of language—where the learner has to actively construct correct sentences on a continual basis—rather than passively absorbing grammatical rules, lexicons, etc. While passive absorption of knowledge may be appropriate at the elementary level, it is not helpful at the intermediate-to-advanced levels, where the learner should learn to produce correct language forms in varied complex contexts.

Technical features

Some of the technical aspects of our approach were introduced previously, (Katinskaia et al., 2017). At present, the system works with several endangered minority languages (listed with their respective language families):

- Komi-Zyrian (minFU)
- Udmurt (minFU)
- Meadow Mari (minFU)
- Erzya (minFU)
- Sakha (minTU)

We have also begun work on North Saami, and several other languages. The system also works with several “major” languages. This was initially motivated by technical necessities. Most of the minFU languages are inside Russia, where Russian is the dominant language. We observed that a large proportion of texts contain spontaneous code-switches into Russian. This is a normal (socio-)linguistic phenomenon, and must be handled correctly by the learning system—since learning takes place with arbitrary texts. Thus, when processing a text in a minFU language, the computational tools need to be able to analyze Russian as well, to detect the code-switches.

As side-effect of our work on code-switching, we undertook work on several “major” languages, including Finnish, Russian, German. Subsequently, work on these languages took its own direction under REVITA. One benefit for the development process is that working on “major” languages allows us to test methods more quickly, since the amount of computational resources, data, tools, etc., is greater for the major languages, which is a practical reality. The methods discussed in this paper are applied to both endangered/less-resourced and the “major” languages.

- Language learning is provided by an intelligent tutoring system (ITS), that supports the student (and the teacher). The ITS is a “virtual tutor,” which aims to simulate some of the functions of a good teacher.
- Since the system is automated, it can scale to any number of students—by contrast, in a classroom setting, the number of teachers needed grows with the number of students.
- REVITA offers practice and exercises based on the texts chosen by the learners—to maximize their interest in the subject. (Texts can also be shared via the platform by friends or teachers.)
- REVITA is intended for intermediate-to-advanced learners—not for beginners. This includes “heritage” learners of endangered languages—who have a passive knowledge from hearing grand/parents speaking the language, but whose own active fluency is lacking.
- The learner uploads any arbitrary text of her choice into the system. The learner can point REVITA to any Web page

---

2 Arbitrary learning texts can be shared between learners via the platform, without violating copyrights.

3 The system is at revita.cs.helsinki.fi

4 MinFU refers to minority Finno-Ugric languages, MinTU—to minority Turkic languages.

5 We apply the term “dominant” to the majority language, having official state status. Speakers of endangered languages are often proficient in the dominant language, especially younger speakers.

6 The same is needed for N. Saami texts which may code-switch into Finnish, etc.
REVITA analyzes the story, sentence by sentence, and automatically generates various exercises—“puzzles”—based on the text. The simplest type of exercise is a “cloze” quiz; a prototypical exercise may work like this: REVITA hides one word in a sentence, and gives the learner a “hint” for the missing word—namely, it’s “base” form. The learner must infer the base form correctly to fit the surrounding context in the story. A snapshot of the practice mode is shown in Figure 1. For some words, multiple-choice quizzes are provided, where the learner must select the correct option from a menu—the most appropriate one, in the context of the story. All exercises are generated from the story automatically.

As the learner answers exercises, the system provides immediate feedback about which answers were correct/incorrect.

The learner can click on any unfamiliar word, to ask the system for a definition or translation into a language of the learner’s choice. All unfamiliar words are stored into decks of flashcards, which the learner can later review—for timed-repetition exercises, to grow the vocabulary.

The system can build random crossword puzzles based on the story. The clues are translations of words hidden from the story. This exercise mode is an additional means of making practice sessions more entertaining, varied and challenging.

The same story may be practiced multiple times. All exercises—clozes, multiple-choice quizzes, crosswords, flashcards, etc.—will be different on each pass through the story, since they are generated randomly on each iteration—to reduce boredom, and keep the learning process fresh, interesting and challenging.

A learner can compete with a friend—play to see who can complete all exercises in the story in a shorter time and with fewer mistakes.

The system remembers the History—all correct and incorrect answers, in all exercises. Based on the history it performs assessment of the learner’s competency. Assessment is needed to determine the best exercises to offer to the learner next, to reduce emphasis on concepts that the learner already has mastered, and those that are too difficult for her at present; see below in section 3.2..

Model-based approach

The achieve the goal of simulating a good teacher, the language learning system should:

1. know the subject matter,
2. know the student: be able to assess the level of competency of the learner,
3. provide personalized instruction: know what skills the learner is best prepared to learn next, to offer exercises that target those skills.

These three functions are accomplished by the Domain model, the Student model and the Instruction model, respectively—see Figure 2.

The Domain model contains representations of the linguistic concepts, which the learner must master for the given language. Concepts do not exist in isolation, but are related in various ways. One important relation is the prerequisite relation. If A is a prerequisite for B, then most students are unlikely to master B before they have mastered A. Relations among concepts are used by the Instruction model.
The Student model identifies each learner’s mastery of concepts over time. For example, in Knowledge Space Theory (Doignon & Falmagne, 2012; Falmagne & Doignon, 2010; Falmagne et al., 1990), the Domain model is represented as the knowledge space—the space of feasible knowledge states of any learner. Each knowledge state is a set of linguistic concepts/skills. A knowledge state S is feasible if it does not contain inconsistent information: if A is a prerequisite of B, and B is in S, then A is also in S. The space is a directed acyclic graph, linking the possible knowledge states.

Ideally, for each given learner, the Student model should specify a probability distribution over the knowledge space—assign a probability that the learner is in the given state in the knowledge space.

The Student model draws on the Domain model and the learner History. The history stores all answers for all exercises. Ideally each exercise should be linked to one (or more) concept. Hence, from the counts of correct/incorrect answers to exercises, we can estimate the probability that the learner has mastered each concept. If the learner’s history contains many entries—many hundreds for each concept—then we can collect reliable statistics, and have accurate estimates of the learner’s competency. Typically the learner’s information is incomplete, and needs to be inferred indirectly. The Domain model contains information about concept dependencies, which helps fill in the gaps in the system’s estimates of the learner’s competency.

The Instruction model builds upon the Domain and the Student model to decide in which order the concepts should be presented in exercises. The key objective and research challenge is accurate assessment of the learner’s competency. Assessment is essential for simulating a good teacher/tutor: if exercises are be too difficult too often, the learner will become discouraged and will quit learning; if exercises are too simple too often, the learner will become bored and will

Leveraging NLP and crowd-sourcing

We emphasized at the outset that the system crucially depends on pre-existing resources for the target language. Such resources are the result of massive documentation and NLP efforts. Next we examine the kinds of resources used for e-Learning, which enable us to build the learning platform. Without such resources, the functionality described above cannot be implemented.

- **Morphological analysis**: a building block on which REVITA rests for many languages is a morphological analyzer. Uralic and Turkic languages have very rich morphology, as do the supported “major” Indo-European languages. Morphological analysis is used for creating exercises. The methodology for building morphological analyzers is well-understood, and has been standard practice for decades, see, e.g., (Koskenniemi, 1983).\(^\text{10}\)

- **Corpora**: the existence of sources from which the learner

\(^{10}\) For many languages (Uralic, Turkic, etc.), we use analyzers developed over the last 10–15 years, in particular, from the Giel-lateko (Moshagen et al., 2013) and the Apertium (Forcada et al., 2011) platforms.
can select interesting content is essential for success. The sources can be static (collections of texts), or dynamic (with changing content, such as on-line newspaper, journals, etc.) For all languages, REVITA provides a “starter” library—a sample of publicly visible texts,\textsuperscript{11} and a list of suggested Web sites, with usable content. In the absence of such corpora, the learner would have to find materials for practice.

- **Bilingual dictionaries:** REVITA provides translations for words unfamiliar to the learner, and the flashcard exercise mode. Without bilingual dictionaries this is not possible.
- **Mono-lingual dictionaries** enable lexicon learning, where for unfamiliar words, REVITA presents the definition in the target language (rather than translations into a different language). This approach is more immersive: the learner does not switch between the target language and the language of translation, working completely within the target language during practice.
- **Machine translation** (MT) models, if available, provide an essential pillar in e-Learning. This is discussed below in Section 4.1...
- **Text-to-speech** (TTS): when possible, REVITA allows the learner to hear how any selected word in the text sounds in context. Further, auditory clues—also produced by the TTS models—are given in listening exercises, in place of base forms or written definitions.\textsuperscript{12}
- **Speech-to-text** (STT): where STT is available, REVITA can accept input from the learner in spoken form. This enables us to give the learner an opportunity to practice speaking (rather than writing only), to provide feedback about the quality of pronunciation, etc.
- **Disambiguation models:** text generally contains a great deal of ambiguity: morphological, semantic, etc. Without a disambiguation model, when the learner clicks on an unfamiliar word in text, which happens to be ambiguous, REVITA provides all of the word’s possible meanings. Ideally, a disambiguation model identifies among all possible analyses the correct analysis, based on the context of the word within the story. At present, most low-resource languages lack disambiguation models; disambiguation models require large training data.
- **Name recognition** models: can provide additional benefits, to make the results of morphological analysis more usable. Certain words may be ambiguous, e.g., between common vs.
  proper noun. If we have a model for identifying proper nouns, REVITA can create exercises for names as well; otherwise names are not used in creating exercises.
- **Test materials:** for some languages test materials are available—questions prepared by teachers/language experts for assessment of competency across linguistic skills. REVITA supports timed testing sessions, with feedback to the learner about her performance levels, broken down by category, and tracked across time.\textsuperscript{13}

Test sets are a valuable resource for assessment. For example, for Finnish and Russian, REVITA uses a database of thousands of questions compiled by teachers over two decades, designed to test fine-grained linguistic “concepts,” from various categories: nominal and verbal inflection, case government, collocations, syntax, orthography, etc.

### Machine translation in e-Learning

The user needs translation into her native/preferred language when faced with complex phenomena. MT models are typically available for major languages, which have large training resources, such as bi-lingual/parallel corpora. The quality of translation directly depends on the amount of available training data. Despite major advances in MT, by means of recent neural network models, e.g., (Vaswani et al., 2017a; Bahdanau et al., 2015), high-quality translation remains an elusive goal.

Translation of sub-par quality is not usable in an e-Learning environment. We must avoid giving the learner incorrect or misleading information at all costs: once the learner loses trust in the “tutor,” she will abandon learning. Thus, as a general principle in e-Learning, when faced with a choice between an error of commission vs. of omission, we must err on the side of omission. If the system is uncertain, it is always safer to give the user no information, than to give information that is wrong.

On the other hand, in the e-Learning environment we do not require full-sentence translation—especially at the LIA levels. Translating smaller linguistic units is sufficient in many cases.

Translation at word level is achieved through dictionaries, discussed above. If we can provide translation of multi-word expressions (MWE), much of the translation needs will be satisfied. Translating MWEs, smaller phrases, is a simpler problem than full-sentence translation, requiring fewer resources.

Identifying non-compositional expressions\textsuperscript{14} is researched extensively, including in our earlier work, e.g. (Kopotev et al., 2013). Combined with “light-weight” translation models, these cover a large portion of the translation needs in e-Learning.

For the less-resourced languages, several research directions aim to compensate for the scarcity of parallel corpora. One direction is on learning to transfer translation models from larger but related languages, (Sharoff, 2018); another is on translation models from “comparable” corpora (rather than parallel corpora, which are more difficult to obtain), or even directly from different mono-lingual corpora (Conneau et al., 2017). As these models evolve and improve, they will be directly usable in e-Learning for less-resourced languages.

### Sequence-to-sequence models

In this section we explore how state-of-the-art methods that are used for translation can be applied more broadly in the context of language learning. The most successful translation models at present are based on neural network (NN) architectures. Important examples are recurrent neural networks (RNN), including long short-term memory networks

---

\textsuperscript{11} We have obtained permission from the copyright holders.

\textsuperscript{12} Building TTS models requires substantial amounts of annotated audio data, which is far from trivial to collect. Currently TTS models are available only for major languages.

\textsuperscript{13} Currently implemented for several major languages.

\textsuperscript{14} The meaning of a non-compositional MWE is not composed of the meanings of its parts—the meaning of “hot dog” in English is not combined from the meanings of “hot” and “dog”.
Leveraging crowd-sourcing

We have described features of REVITA that help the learner to improve his or her level of competency. We next explore the converse question—i.e., how the quality of these features can improve over time, automatically, from the input that we can gather from large learner communities. REVITA provides a wide range of crowd-sourcing possibilities. One straightforward application of crowd-sourcing is automatic improvement of underlying linguistic resources. For example, freely-available bi-lingual dictionaries may contain inaccuracies. REVITA users can edit their flashcards. If we observe that A, many learners add (or remove) a translation for a given word, and B, that these learners are “trustworthy” (e.g., they perform consistently well on exercises, etc.), then we may infer that the correction suggested by these learners is valid, and enhance the resource.

An important crowd-sourcing possibility is using the platform for creation of learner corpora. A learner corpus is an especially valuable resource, which is difficult to collect, and which offers a view of the kinds of mistakes that learners are likely to make at different stages of learning. From learner corpora researchers collect statistics about which mistakes are more common under which conditions; as always, the larger the learner corpus, the more reliable the statistics.

For each endangered language that the learners study, REVITA collects all answers—correct and incorrect—into the History. The History drives the selection of exercises, as described in Section 3.1.; the History also constitutes a learner corpus.

As an example of the benefits that we can obtain from such a learner corpus, consider the problem of generating exercises. As shown in Section 3., we generate cloze exercises from any story (see Figure 1). We plan to adapt REVITA for mobile devices, to enable learners to engage with the platform more often, without a need for a desktop or laptop computer. Cloze quizzes require typing on a keyboard, which is less convenient on a mobile device. Hence, on mobile devices it is desirable to replace cloze quizzes by multiple-choice quizzes. In this case, if the system can refer to a massive learner corpus to find what mistakes learners typically make in similar contexts, it can offer these mistakes as distractors—options given to the learner in a multiple-choice menu. This will facilitate practice on mobile devices.

To obtain high-quality distractors, large amounts of learner data need to be collected. The main point is that without a platform such as REVITA, such data is impossible to collect in sufficient amounts.

Crowd-sourcing requires the participation of many learners, and, over time, yields many benefits. We do not go deeper into crowd-sourcing in this paper; this is the focus of much research in the emerging field of educational data science, (Romero & Ventura, 2010).

Learning Domain models from data

Each exercise can be linked to a linguistic “concept” (also called “skill” in Knowledge Space Theory (KST)). An example of a concept may be case government—a certain set of verbs or prepositions require their nominal object to be in a particular case. Creating a list of the concepts that make up the domain model needs to be performed by domain experts, for each language.

For each language, we may have some hundreds of concepts. Concepts are not learned independently. The student will find it natural to learn them in some order—because some concept is a prerequisite for another, or because some advanced concept is best left for later, when the student has a higher level of proficiency.17 We next describe a simple baseline model to infer such relations from crowd-sourced user data.

15 These models follow the encoder-decoder approach, where the source string is first “encoded” into an object which stores all the information necessary for a following decoder phase, where the target string is generated.

16 These are not yet fully implemented in REVITA, but are rather work in progress.

17 When one concept precedes another one, it is called the surmise relation in KST.
Baseline approach to learning Domain models Each exercise is linked to a certain concept, and we store all exercise results from every student. From these results, we can tell which users “know” which concepts, and thus we can try to tell apart the more basic concepts from the more advanced ones.

We aim to build a partial order over the set of all concepts $C$, which will give us the pairs of concepts that are related—we write $c_2 \rightarrow c_1$ to mean concept $c_2$ presupposes concept $c_1$.\footnote{We would “like” to say $c_1$ is a prerequisite for $c_2$, but that claim would be too strong. However, we can conclude from the data that typically $c_1$ is mastered before $c_2$.}

Given a set of users $U$, we build a matrix of “mastery” scores $M$, of dimension $|U| \times |C|$, where each element $(i, j)$ is the proportion of correct answers of user $u_i$ for concept $c_j$ that were correct.

We can apply a threshold $\theta_{\text{know}}$ to each element of $M$, to decide whether the user “knows” that concept, and obtain a binary matrix $M'$:

$$M'_{ij} = \begin{cases} 1 & M_{ij} > \theta_{\text{know}} \\ 0 & \text{otherwise} \end{cases}$$

Next, for every pair $c_1, c_2 \in C$, we determine if $c_2 \rightarrow c_1$, or $c_1 \rightarrow c_2$, or neither. For this, we pick all possible column pairs in $M'$ and apply the logical implication operation row-wise:

<table>
<thead>
<tr>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$c_2 \rightarrow c_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

For each pair of concepts $c_1, c_2$, this gives us a binary vector $x_{c_1c_2}$ of length $|U|$. Each element of $x_{c_1c_2}$ corresponds to a user in $U$, and tells whether this user’s state of knowledge is consistent (1) or inconsistent (0) with the assumption that $c_2 \rightarrow c_1$. We compute the average value of its elements, $\overline{x_{c_1c_2}}$, and apply another threshold, $\theta_{\text{cons}}$—a consistency threshold: if $\overline{x_{c_1c_2}} > \theta_{\text{cons}}$, then we add the relation $c_2 \rightarrow c_1$ to our partial order.

Finally, we represent our partial order as a directed acyclic graph, where each path represents a possible route to learning a concept. For example, if we wish to obtain a complete syllabus for a language course, we can find a total order compatible with our partial order (i.e., a linear extension) by topologically sorting the nodes in the graph.

We tested this approach with a set of over 150,000 answers gathered from 400 users, and manually evaluated the results, with $\theta_{\text{know}} = .90$, and $\theta_{\text{cons}} = .95$. Domain experts confirmed that the graph provides a correct model for the relations between the concepts.

Conclusion and current work

Two factors complicate formal evaluations of our proposed approach. First, in this stage of the work, it is difficult to obtain measurable quantitative results in terms of effectiveness and impacts. Automated personalized tutoring tools do not exist even for major languages,\footnote{NB: we consider here as relevant only tools at the intermediate-to-advanced levels, rather than at the elementary levels. At the lower levels, there is an abundance of on-line learning tools for many languages.} much less for endangered ones. A detailed survey that we conducted, (Katsinka et al., 2018), as well as other surveys, confirm that the currently available CALL systems do not exhibit sufficiently intelligent characteristics, and ITS—intelligent tutoring systems\footnote{For tutoring in general, not necessarily in languages.} described in the literature appear to be laboratory experiments, and are not available for use by learners in practice.

At present, to the best of our knowledge, REVITA is the only platform for learning/tutoring at the intermediate-to-advanced levels, that is functional, freely-available and supports multiple languages, including endangered languages.\footnote{One approach introduced by the VIEW/WEITi (Working with English Real Texts) system, (Meurers et al., 2010), is in some respects related to our methodology. It supported up to four “major” languages, no endangered ones.}

Second, for extra-linguistic reasons, it has proven non-trivial to engage learner communities of minFU languages in using the e-Learning tools. However, our experience with several major languages—at present, Finnish, Russian, German, and Kazakh—has been quite positive. We have received excellent feedback from language teachers in over 10 universities, and hundreds of intermediate-to-advanced students are finding REVITA to be a beneficial resource for their studies. This gives us hope that in near future, we will be able to collect reliable statistics that will enable us to measure the effectiveness of the platform for endangered and less-resourced languages.

In conclusion, we note that all of the fundamental linguistic resources on which REVITA rests, listed in Section 4, are abundant for the major languages. For the less-resourced languages, building each resource requires a substantial investment of human documentation effort. Our main claim is: while each of the listed resources in isolation may do little to effect revitalization, combined with our proposed approach, they yield a powerful learning tool, much greater than the sum of its parts.

While our approach may not be feasible for all endangered languages, we try to show how the technology-aided tools may provide a chance to revitalize some of the endangered languages. Thus we propose that REVITA points the way to combining valuable resources, collected through research and documentation, into novel tools that can actually help the speaker communities of endangered languages to sustain their language and culture.

Acknowledgements

We thank J.M. Hoya Quecedo, S. Ivanova, A. Katsinkaia, M. Koppatz, and J. Nouri, and colleagues at other universities for their contributions. This work was supported in part by Project “FinUgRevita,” Academy of Finland, grant No. 267097.

Bibliographical References


