Tools for supporting language learning for Sakha

Sardana Ivanova, Anisia Katinskaia, Roman Yangarber
University of Helsinki
first.last@helsinki.fi

Abstract

This paper presents an overview of linguistic resources available for the Sakha language, and presents new tools for supporting language learning for Sakha. The essential resources include a morphological analyzer, digital dictionaries, and corpora of Sakha texts. We extended an earlier version of the morphological analyzer/transducer, built on the Apertium finite-state platform. The analyzer currently has an adequate level of coverage, between 86% and 89% on two Sakha corpora. Based on these resources, we implement a language-learning environment for Sakha in the Revita computer-assisted language learning (CALL) platform. Revita is a freely available online language learning platform for learners beyond the beginner level. We describe the tools for Sakha currently integrated into the Revita platform. To our knowledge, at present this is the first large-scale project undertaken to support intermediate-advanced learners of a minority Siberian language.

1 Introduction

The Sakha language, also known by its exonym *Yakut*, is the language of an ethnic community, who mainly inhabit the Republic of Sakha in the Far East of Siberia, Russian Federation. According to the 2010 census, Sakha is the native language of 450,140 people, and is considered vulnerable due to its limited usage. Children do not use Sakha in all aspects of their life; they speak Sakha at home with family, but do not use it in school and socially.

Sakha belongs to the Northern group of the Siberian branch of the Turkic language family, and is agglutinative, as are all Turkic languages, (Ubryatova, 1982). It has complex, four-way vowel harmony, and a basic Subject-Object-Verb word order. The lexicon of Sakha consists of native Turkic words, has many borrowings from the surrounding Mongolic and Tungusic languages, numerous loanwords from Russian, as well as words of unknown origin. Sakha makes extensive use of post-positions, which indicate syntactic relations and govern the grammatical case of nominals, (Forsyth, 1994).

In the digital sphere, Sakha can be considered a low-resource language. We report on our project to provide learning support for Sakha. Building on pre-existing digital resources, we aim to provide a learning platform for students (including adults) who are interested in strengthening their linguistic competency in Sakha.

The paper is structured as follows. Section 2 describes distinctive properties of the Sakha language and motivates the need for language-learning support by reviewing the social environment of the language. Section 3.1 presents an overview of previous work Sakha; Section 3.2 describes the Revita platform for language learning. Section 4 describes the instruments we integrate to support language learning for Sakha. In Section 5 we discuss initial results obtained with the tools. Sections 6 concludes with pointers for future work.

2 Sakha language

Sakha is the national language of the Sakha people, which, along with Russian, is one of the official languages of the Republic of Sakha (Yakutia), (Yartseva, 1990). The Sakha language differs significantly from other Turkic languages by the presence of a layer of vocabulary of unclear (possibly Paleo-Asiatic) origin, (Kharitonov, 1987). There are also a large number of words of Mongolic origin related to ancient borrowings, as well as late borrowings from the Russian language, (Tenishev, 1997).

2.1 Distinctive features

Vowels in Sakha follow complex vowel harmony rules. The features of the vowels within a word must agree in a strictly defined fashion. First, *palatal-*

velar harmony of vowels is observed in Sakha strictly sequentially and does not admit exceptions. If the first syllable contains a front vowel, then the vowels in all subsequent syllables in the word must be front. Otherwise, if the first syllable contains a back vowel, then the vowels in all subsequent syllables must be back. Second, *labial* vowel harmony requires that the sequence of vowels agree according to the degree of roundedness within adjacent syllables, (Sleptsov, 2018). For example:

- back+unrounded: "аҕа / аҕалардыын"
  [aɣa / aɣalardīn]
  "father/with fathers"
- back+rounded: "оҕо / оҕолордуун"
  [oɣo / oɣolordūn]
  "child / with children"
- front+unrounded: "эбэ / эбэлэрдиин"
  [ebe / ebelerdīn]
  "grandmother / with grandmothers"
- front+rounded: "бөрө / бөрөлөрдүүн"
  [börö / börölördǖn]
  "wolf / with wolves"

Thus, the vowels in the suffixes "-лар-" [lar], indicating the plural, and "-дыын" [dīn], indicating comitative case, undergo 4-way mutation according to vowel harmony.

In Sakha, the verb is the central part of speech, (Dyachkovsky et al., 2018). Some verbs can have multiple affixes (as in most Turkic languages), which can correspond to an entire clause or sentence in other languages, such as Russian. Sakha has no infinitive form for verbs, therefore a predicate that (in other languages) would include an infinitive is conveyed by various indirect means, for example:

- "суруйан бүтэрдэ": [surujan büterde]
  "he finished writing"
  (literally: *"he wrote, finished"*);
- "сатаан ыллыыр": [satān ıllīr]
  "he can sing"
  (literally: *"he knows how, sings"*);
- "бобуоххун син": [bobuoχχun sin]
  "you can forbid"
  (literally: *"you can, let's forbid"*).

Sakha is characterized by an exceptional variety of verbal tenses. In particular, according to (Korkina, 1970), 8 past forms are distinguished:

- proximal-past:
  "үлэлээтим" [ülelētim]
  "I worked (recently)";
- remote-past:
  "үлэлээбитим" [ülelēbitim]
  "I worked (long ago)";
- past perfect:
  "үлэлээбиппин" [ülelēbippin]
  "In fact, I worked";
- episodic past:
  "үлэлээбиттээхпин" [ülelēbittēχpin]
  "I used to work on occasion";
- past imperfect:
  "үлэлиирим" [ülelīrim]
  "I worked in the past for some time";
- plusquamperfect:
  "үлэлээбит этим" [ülelēbit etim]
  "I had worked prior to that";
- episodic plusquamperfect:
  "үлэлээбиттээх этим" [ülelēbittēχ etim]
  "Long ago, I used to work".

The total number of tense forms exceeds 20.

One of the particularities of nouns is when *paired nouns* are marked with possessiveness, both components of the compound noun change *in parallel*, as the word is inflected:

- "баай-дуол" [bāj duol] *"wealth"*:
  "баайа-дуола" [bāja duola]
  (3rd person possessive, nominative case),
  "баайын-дуолун" [bājın duolun]
  (3rd person singular possessive, accusative)
- "сурук-бичик" [suruk bičik] *"writing"*:
  "сурукта-бичиктэ" [surukta bičikte]
  (partitive case),
  "суругу-бичиги" [surugu bičigi]
  (accusative case)

2.2 Socio-linguistic environment

According to (Vasilieva et al., 2013), since 1990, the percentage of ethnic Sakha has grown, reaching 45% of the total population in the Republic of Sakha. Ethnic Sakha together with other indigenous peoples of the North Siberia and the Far East comprise over 50% of the total population.

Vasilieva et al. (2013) has conducted surveys, which show a direct dependence of the level of linguistic proficiency on the *language of instruction* at school. A fluent level of proficiency is achieved by:

- respondents who had schooling in the Sakha language (34.5%)

- respondents who had studied in schools, where subjects were taught in Russian and partly in Sakha (27.4%).

Only 17.9% of respondents who had studied in Russian are fluent in Sakha. Respondents who speak Sakha poorly, or do not speak at all, graduated from Russian-speaking schools. Thus, as expected, linguistic skills and abilities in Sakha are poorer for those who had studied in Russian.

In work life, the Russian language is dominant for all age groups. In the two youngest age groups (16–25 and 26–35 years old), the use of Russian is growing, approaching 50%. This is due to the requirements of formal communication, terminological dependence, ethnically mixed composition of professional teams. On the other hand, after the completion of active professional life, the return to an increased usage of the original ethnic language is common, (Vasilieva et al., 2013).

In Yakutsk—the capital and the largest city of the Sakha Republic—one in three Sakha children lack the opportunity to study in their native language. This is a violation of the right to study in one's native language. The number of schools which offer teaching in Sakha in Yakutsk in 2002–2003 was 16, and dropped to 15 by 2003–2004. The number of schools where Sakha is studied as a subject decreased from 22 (in 2002–2003) to 16 (in 2003–2004). The number of Sakha language learners decreased from 6,377 to 2,902, (Vasilieva et al., 2013). According to the statistical report of the Ministry of Education of the Republic of Sakha in 2006–2007, the *cities* of the Republic had 147 schools with Russian language of instruction (61,055 children), 4 educational institutions with non-Russian languages of instruction (1014 children), 29 institutions with a mix of Russian and non-Russian languages of instruction (18,094 children). In 11 schools (serving 1,262 students) non-Russian languages are offered as optional subjects of study.

Vasilieva et al. (2013) indicate that this situation concerning the language of instruction of ethnic Sakha pupils has a direct correlation with other serious problems in terms of linguistic competency in Sakha and vitality of Sakha—acculturation and assimilation of urban youth, which will leads to linguistic conformism due to the lack of sufficient social opportunities for using the language.

3 Prior work

3.1 Sakha language resources

Despite the current advances in digitization, digital resources for the Sakha language are severely lacking. The creation of digital tools would strengthen the language in a number of ways, and several projects are being undertaken to support Sakha. We briefly mention some of them here.

The digital bilingual dictionary `SakhaTyla.Ru`[1] currently offers over 20,000 items from Sakha to Russian, over 35,000 items from Russian to Sakha, about 2,000 items from Sakha to English, and about 1,000 items from English to Sakha. In addition to translations, this dictionary also contains *examples of usage*, including idiomatic usage, for every item, which constitutes a base of lexical data, and can be highly useful for language learning and teaching. The base of examples from this dictionary is currently not utilized in our learning platform.

Leontiev (2015) has compiled a newspaper corpus of Sakha containing over 12 million tokens. The Sakha Wikipedia contains over 12,000 articles, which makes up a corpus of over 2 million tokens.[2]

A Sakha course on the educational platform Memrise offers a vocabulary of about 3100 words.[3]

Audio materials: Common Voice is a platform for crowdsourcing open-source audio datasets.[4] At present, it offers just under 2.5 hours validated voice recordings in Sakha. By comparison, English has almost 850 hours of audio content on the platform, and Russian has 50 hours.

In summary, few linguistic resources exists for Sakha.

3.2 Revita language learning platform

Revita is an e-learning platform, which uses methods from computer-assisted language learning (CALL) and intelligent tutoring systems (ITS).[5] The platform provides a language-independent foundation for language learning, which can be adapted to support new languages, by adding language-specific resources, without modifying the core system. The platform is used for language teaching and learning at several universities in Europe and Asia.

The goal of the system is to provide tools for language learning, (Katinskaia et al., 2018), and to support endangered languages, (Katinskaia and Yangarber, 2018; Yangarber, 2018). The system focuses on stimulating the student to actively produce language, rather than passively absorb exam-

[1] www.sakhatyla.ru

[2] sah.wikipedia.org

[3] www.memrise.com/course/153579/sakha-tylyn-leksikata-sakha-tyla-iakutskii/

[4] voice.mozilla.org/en/about

[5] revita.cs.helsinki.fi

ples of language use or grammatical rules. The system achieves this by helping students learn language from *stories*. The story can be any text, which the students can choose themselves. The platform takes an arbitrary text chosen by the user and uses it as practice material; it creates exercises for the student based on the text: the exercises are new every time the student practices with the text—to keep the practice sessions interesting and to reduce boredom. The computational engines in the platform analyze the text, and try to determine which concepts are best suited for the student to learn next.

The platform has been customized for several less-resourced Finno-Ugric languages: Erzya, Komi-Zyrian, Meadow Mari, North Saami, Udmurt; it has also been customized for Kazakh (also a Turkic language), and several others. Revita also offers a number of languages with larger resources: currently, the most developed are Finnish, Russian, and German, and initial support exists for Swedish, Spanish, Catalan, and French.

4 System for supporting Sakha

In this section we describe our work on adding the Sakha language to the Revita platform for language learning. The system is built on several lower-level linguistic tools and components.

4.1 Morphological analyzer

The morphological analyzer, part of the package called `apertium-sah`, was developed in the context of the Apertium platform, (Forcada et al., 2011). The analyzer is developed using the Helsinki Finite-State Toolkit (HFST), (Lindén et al., 2011). The lexicon and morphotactics are written in the `lexc` formalism, and the morphophonology is developed using the `twol` formalism, based on the Two-Level Morphology framework, (Koskenniemi, 1983). The transducers are compiled into a morphological analyzer and a generator. The transducer is two-directional: on one hand, it can map a surface form to all of its possible lexical forms; on the other hand, it can take a lexical form and generate all of its corresponding surface forms. That is, the transducer can be used both for analysis and for generation of surface forms.

For example, the surface form “атын” receives two analyses—lexical forms:

- ат [at] <n> <px3sg> <acc>
  Horse.Noun.Possessive-3sg.Accusative
  “his horse” (accusative)
- атын [atın] <adj>
  “other” Adjective (indeclinable)

We extended the initial, baseline version of the Apertium Sakha analyzer by adding lemmas to the lexicon based on their frequencies, which we computed from the Wikipedia corpus. Initially the analyzer had 4,303 stems. Table 1 gives the number of lexical items for each of the major parts of speech (POSs) in inital and extended analyzer versions.

Part of speech	**Original**	**Improved**
Noun	2,582	4,240
Proper noun	815	2,155
Adjective	464	1,362
Verb	278	1,038
Adverb	62	338
Numeral	58	89
Pronoun	15	17
Postposition	12	42
Conjunction	7	16
Determiner	10	16
Total:	4,303	9,313

Table 1: Number of stems per part of speech

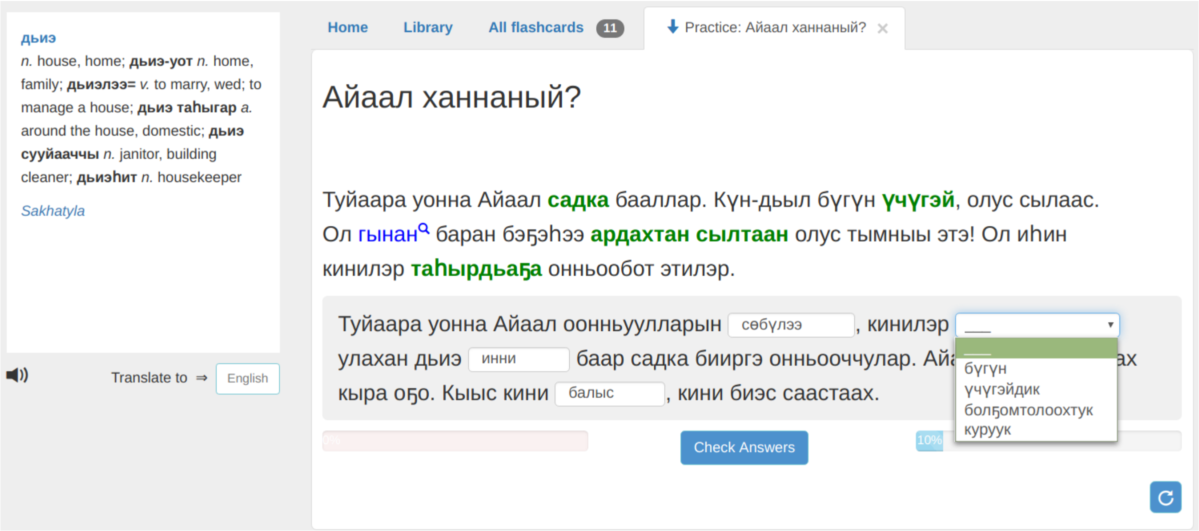
The morphological tagset consists of 92 tags: 16 tags indicate parts of speech—noun, adjective, verb, postposition, etc.—and 76 tags indicate values for morphological subcategories, e.g., for case, number, person, possession, transitivity, tense, aspect, mood, etc. We consulted Ubryatova (1982) as the principal source of grammatical information.

4.2 Language learning platform

The platform offers the learner several exercise modes based on input stories: reading mode, practice mode, flashcards, crossword mode, etc.

In the *reading mode* learner can read a story, and request translations of unfamiliar words.

In the *practice mode* the system generates exercises based on the story, which user has uploaded to the system. The story undergoes several stages of analysis. At the lowest level the system uses the Sakha morphological analyzer, (Ivanova et al., (To appear). The story is presented in “snippets”—small pieces of text, about 50 tokens each, approximately one paragraph or 2–3 sentences in length. The system selects some of the tokens in the snippet to generate quizzes. Each quiz may be of several types: “cloze” (i.e., fill-in-the-blank quiz), multiple-choice, or a listening exercise (where the learner must type in the words s/he hears).



дьиэ
n. house, home; **дьиэ-уот** *n.* home, family; **дьиэлээ=** *v.* to marry, wed; to manage a house; **дьиэ таһыгар** *a.* around the house, domestic; **дьиэ сууйааччы** *n.* janitor, building cleaner; **дьиэһит** *n.* housekeeper

Sakhatyla

Translate to ⇒ English

Айаал ханнаный?

Туйаара уонна Айаал **садка** бааллар. Күн-дьыл бүгүн **үчүгэй**, олус сылаас. Ол гынан баран бэҕэһээ **ардахтан сылтаан** олус тымныы этэ! Ол иһин кинилэр **таһырдьаҕа** онньообот этилэр.

Туйаара уонна Айаал оонньуулларын сөбүлээ , кинилэр ___ улахан дьиэ инни баар садка биирги онньоочулар. Айа... ах кыра оҕо. Кыыс кини балыс , кини биэс саастаах.

Figure 1: Practice mode

The system creates cloze quizzes from *inflected* parts of speech, i.e., nouns, verbs, etc. For example, the first sentence in the snippet in Figure 1 is

> "Туйаара уонна Айаал оонньуулларын сөбүлүүллэр."
> [tujāra uonna ajāl ōnnʲūlların söbülǖller]
> "*Tuyaara and Ayaal like to play.*"

The current snippet of the story appears over a grey background, and contains cloze and multiple-choice exercises. The system created a cloze exercise, showing to the user only the base form (lemma) "сөбүлээ" [söbülē] ("*to like*") of the surface form "сөбүлүүллэр" [söbülǖller] ("*they like*"):

> "Туйаара уонна Айаал оонньуулларын *сөбүлээ*..."
> [tujāra uonna ajāl ōnnʲūlların söbülē]

From the verb lemma the learner should guess which form of the hidden word fits the context best.

Multiple-choice quizzes are constructed also from non-inflected parts of speech (adverbs, postpositions, etc.) Tokens of similar part of speech are presented to the learner as "distractors"—incorrect answer options. Figure 1 shows a multiple-choice quiz for the token "куруук" [kurūk] ("*always*") with other adverbs serving as distractors.

Listening exercises (optional) are generated from tokens in the story—the words are spoken by a speech synthesizer and the learner must enter the word that was pronounced. Currently, listening exercises are not available for Sakha; we plan to incorporate them into the system when text-to-speech (TTS) synthesis for Sakha becomes available.

The previous snippets—above the current snippet—show correctly answered questions—coloured in green—and incorrectly answered questions in blue.

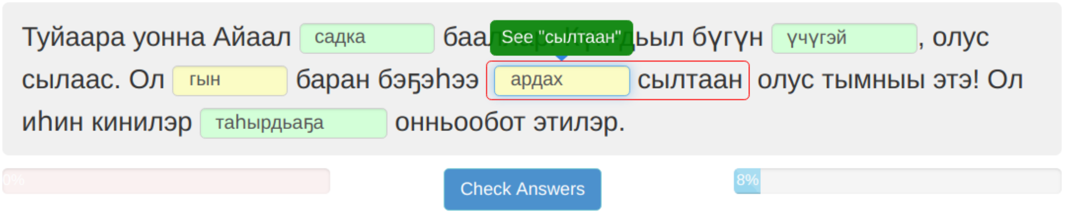
The choice of candidates for exercises depends on the *user model*—based on the history of the user's previous answers. The system computes probabilities (weights) for potential candidates in the snippet. Exercises receive a lower probability if the student had mostly answered them correctly or mostly incorrectly in earlier sessions—since it means that they are too easy or too difficult for the learner at present.

In the *crossword mode*, a crossword is built based on the text. Exercises for the crossword are selected randomly, and according to the same principles as in practice mode.

The user can receive the translation of an unfamiliar word by clicking on it. The box on the left in Figure 1 shows a dictionary entry for a token clicked by the user—"дьиэ" [ǯie] ("*house*"). The learner can request a translation of an unfamiliar word in all practice modes. Translations are looked up in the SakhaTyla.Ru digital dictionary. The system records the words for which translations were requested into the user's own set of flashcards. In the flashcard mode the user can practice vocabulary, using timed repetition algorithms.

Stories for learning can be found on newspaper websites, such as edersaas.ru, kyym.ru, etc., or from the Sakha Wikisource.[6]

[6] https://sah.wikisource.org/



Туйаара уонна Айаал садка баа ьыл бүгүн үчүгэй, олус сылаас. Ол гын баран бэҕэһээ ардах сылтаан олус тымныы этэ! Ол иһин кинилэр таһырдьаҕа онньообот этилэр.

Figure 2: Example of a chunk

4.3 Chunking for exercises

Revita allows the language expert to customize the system for a new language by explicitly specifying rules for syntactic government and agreement. The system performs shallow parsing ("chunking") based on these rules, and uses the chunks when creating exercises, discussed in the previous section. Next we discuss how these rules can work for Sakha.

A government rule, R_g, may state:

(сылтаан, PostP) → [Ablative]

which means that the post-position lemma *сылтаан*, [sıltān], (meaning "because-of"), *governs the ablative* case of its preceding noun (or noun phrase).

A simple but quite general agreement rule, R_a, may state:

[{pos: Noun, case: $X }
+ {pos: PostP, gov_case: $X }]

Rule R_a consists of two elements/tokens, and describes case agreement. If a token with noun POS is followed by a token with postposition POS, they will form a unit (phrase) if the free variable $X, indicating the value of the case feature of the noun and the case that the postposition governs, has the same value for both tokens.

Using these two rules, the systems will match all corresponding constructions in text. For example:

"... ардахтан сылтаан ... "
[ardaχtan sıltān]
Rain.Noun.ABL *because_of*.Post-position
"... because of rain ..."

The Revita language learning system uses these chunk rules to construct exercises. For example, the exercise based on the second sentence in Figure 2:

"Ол гын баран бэҕэһээ ардахт сылтаан олус тымныы этэ!"
[ol gın baran beɣehē ardaχ sıltān olus tımnī ete]
"But it was very cold yesterday because of the rain!"

Corpus	Tokens	% Coverage
Original analyzer:		
Wikipedia 2015	1,020,000	73.02
Kyym (newspaper)	1,040,000	71.36
Improved analyzer:		
Wikipedia 2019	2,195,565	89.28
Newspapers	16,436,999	86.41

Table 2: Coverage of the morphological analyzer

The boxes contain the *cloze* quizzes—exercises for the user. The system provides hints for each cloze. First, inside the box it shows the lemma of the word. Further, each phrase circled in red forms a chunk/unit—based on the government and agreement rules, such as R_g and R_a, above. Thus, the post-position "сылтаан" [sıltān] (*"because of"*), which governs the ablative case, links it to its preceding noun "ардах" [ardaχ] (*"rain"*) to hint to the user that the noun's surface form should be in the *ablative*.

5 Discussion

5.1 Analyzer coverage

Table 2 shows the coverage of the improved morphological analyzer, as compared to the original one. Coverage of the original analyzer was measured on the Wikipedia corpus (dump from 2015) and the "Kyym" Sakha newspaper.

The improved analyzer was tested on a Wikipedia dump from 2019, and the large newspaper corpus compiled by Leontiev (2015). Currently, the coverage on Wikipedia is about nine out of ten tokens, which is higher, as expected, since the analyzer was developed based on a frequency list from this corpus.

5.2 Learner engagement

We have presented the language learning platform to language experts and lecturers at the Department of the Sakha language, at the North-Eastern Federal University in Yakutsk, Russian Federation. The experts confirm that the language learning system can be a promising tool for enhancing instruction in Sakha. We also plan to introduce the learning system to Sakha learners in cooperation with *Yakutia.Team*, the organization for promoting Sakha language and culture.[7]

Releasing the system for use by language learners will yield mutual benefits for the learners as well as for the researchers. The learners receive a training platform which helps them improve their linguistic competency. From the interaction of the learners with the platform, the researchers receive valuable educational data, for modeling the process of learning Sakha, the patterns of mistakes that the learners make over time, and insights into how the learning system can be improved on the basis of the collected data.

5.3 Multiple admissibility

Multiple Admissibility (MA) in language learning occurs when more than one surface form of a given lemma fits syntactically and semantically within a given context. MA implies that multiple alternative answers are "correct" for the given context, not only the word that the author chose to use in the story. From the perspective of CALL and ITS (intelligent tutoring systems), MA forms a complex challenge, discussed in current research, (Katinskaia et al., 2019).

The Sakha language presents a particularly rich source of scenarios for multiple-admissible answers. Due to the agglutinative morphology of Sakha, the learner can add affixes to a word, which carry additional information or connotations to slightly alter the meaning of the word. We briefly discuss several such scenarios.

The category of *possessiveness*—possessive affixes on nominals—is one of the fundamental categories of Sakha grammar. Possessive forms are very common, and the scope of their usage is far wider than merely indicating possession in the strict sense; possessive affixes express a wide range of logical relations and connections between objects, which are often *not* related directly to the strict notion of possession, (Ubryatova, 1982). For example:

- "сылдьыбатах сирим"
  [sılǯıbataχ sirim]
  "a place where I have not been"
  (literally: "my place, where ...")
- "билбэт киһитэ"
  [bilbet kihite]
  "man whom s/he doesn't know"
  (literally: "her man, whom ...").

We often find examples of using an impersonal form of a noun in place of a possessive form of a noun and vice versa, as in the following examples:

- Example of using an impersonal noun instead of a possessive noun:
  1a. Form which was used in a story:
  "сааны ылан сүгэр"
  [sānı ılan süger]
  (*"he hung a gun on his shoulder"*)
  1b. The learner's input:
  "саатын ылан сүгэр"
  [sātın ılan süger]
  (*"he hung his gun on his shoulder"*)
- Example of using a possessive form instead of an impersonal form:
  2a. Form which was used in a story:
  "мооньнун уһатан уутун көрөр"
  [mōnnʲun uhatan ūtun körör]
  (*"craning his neck, he looks at (his) water"*)
  2b. The learner's input:
  "мооньнун уһатан ууну көрөр"
  [mōnnʲun uhatan ūnu körör]
  (*"craning his neck, he looks at the water"*)

Secondly, Sakha has a highly developed system of verbal *aspects*. Aspect in Sakha can be expressed by various affixes or analytically. Aspect is one of the most commonly used grammatical categories in Sakha, which allows statements in the language to be expressive and precise, (Ubryatova, 1982). Eight forms are used to designate actions that have occurred prior to the present time.

As a result of this choice, it may be quite difficult to decide which form best fits the context, given only the base form of the word. For instance, it can be difficult to distinguish the "first past" perfect tense and recent past tense, because the results of both actions are connected with a present moment.

The verb form which was used in a story: first past perfect tense, indicative mood, third person, singular:

> "Эһэм сиргэ сылдьан өрүү түргэн хаамыылаах, оттон бу сырыыга өссө чэпчэкитик үктэнэргэ дылы буолбут."

[7] https://www.yakutia.team

[ehem sirge sılǯan örǖ türgen χāmīlāχ, otton bu sırīga össö čepčekitik üktenerge dılı $\boxed{\text{buolbut}}$]
“Grandfather usually walks fast in the field, but this time he also seems to step lighter” (meaning, the action is inferred from its result).

The learner’s input was a verb form in the recent past tense, indicative mood, third person, singular:

“Эһэм сиргэ сылдьан өрүү түргэн хаамыылаах, оттон бу сырыыга өссө чэпчэкитик үктэнэргэ дылы $\boxed{\text{буолла}}$.”
[ehem sirge sılǯan örǖ türgen χāmīlāχ, otton bu sırīga össö čepčekitik üktenerge dılı $\boxed{\text{buolla}}$]
“Grandfather usually walks fast in the field, but this time he also steps lighter” (meaning, the action is observed by the speaker).

These and many other examples show that the task of generating exercises automatically for Sakha is far from trivial and requires much research due to extensive multiple admissibility.

6 Conclusion and future work

This paper offers an overview of the resources available for Sakha, and describes our work on creating tools to support language learning for Sakha. Our surveys of available resources demonstrate that they are severely lacking.

We present the following tools and resources, which we combine to create a system to support language learning for Sakha:

- morphological analyzer, built on the Apertium platform,
- language learning system, built on the Revita platform,
- bilingual Sakha dictionaries,
- several Sakha corpora.

The morphological analyzer is an essential component in any natural language processing (NLP) system, without which little can be done to provide computational tools for the language.

The functionality of the language learning system is under development. For larger languages many more linguistic resources and tools are available than for Sakha. For example, currently, the Sakha system has only noun–postposition government rules. We plan to implement additional shallow-parsing rules to provide intelligent error feedback to the learners. For example, *verb–complement government*:

“ангинанан ыалдьыбыт”
[anginanan ıalǯıbıt]
tonsillitis.Noun.INS *contract*.Verb.2SG.PAST
(“*s|he contracted tonsillitis*”).

The verb “ыарый” [ıarıj] (“*to contract*”) governs the *instrumental* case of the noun “ангина” [angina] (“*tonsillitis*”).

Currently, the system employs chunking (shallow parsing) to track instances of syntactic government. The system can track more complex and longer-range government with the help of deep parsing. Once a parser for Sakha becomes available, it will enable Revita to provide richer feedback to the learner.

The Sakha analyzer needs further improvement to reach higher coverage. The main work to be done is extending the lexicon. While a good level of coverage has been achieved with only 9,313 stems, production-level morphological analyzers have at least tens of thousands of stems—more typically, hundreds of thousands. Once good coverage has been achieved for the morphological analyzer, the next step is to build models for morphological and syntactic disambiguation.

As more advanced tools for Sakha become available, they will be incorporated into Revita, to provide richer functionality:

- Parsers, e.g., dependency parsers, such as based on the constraint-grammar formalism (commonly used in Apertium), or statistical or neural-network based parsers.
- Difficulty models—to predict the difficulty of a story for the user, and to assess the learner’s level of competency—based on how well a learner handles easy vs. difficult stories.
- Disambiguation models—to disambiguate ambiguous tokens in text.
- Text-to-speech—to provide listening exercises based on text.
- Speech-to-text—to provide speaking exercises (not yet available).

Acknowledgements

This research was supported in part by the “FinUgRevita” Project, Academy of Finland, Grant 267097, and EDUFI Fellowship TM-18-10846.

We are very grateful to Javad Nouri, Kim Salmi, Max Koppatz, José María Hoya Quecedo, and Jue Hou for their contributions.

References

Katja Bang. 2015. Language situation in the republic of Sakha (Yakutia). Master's thesis, University of Turku, School of Languages and Translation Studies.

Nikolay Dyachkovsky, Petr Sleptsov, K Fedorov, and M Cherosov. 2018. *Поговорим по-якутски. Самоучитель языка саха (Let's speak Sakha. Sakha language tutorial).* Yakutsk: Bichik.

Mikel L Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M Tyers. 2011. Apertium: a free, open-source platform for rule-based machine translation. *J. Machine Translation*, 25(2).

James Forsyth. 1994. *A history of the peoples of Siberia: Russia's North Asian colony 1581-1990.* Cambridge University Press.

Sardana Ivanova, Francis Tyers, and Jonathan Washington. (To appear). The Apertium implementation of finite-state morphological analysis for the Sakha language.

Anisia Katinskaia, Sardana Ivanova, and Roman Yangarber. 2019. Multiple admissibility: Judging grammaticality using unlabeled data in language learning. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 12–22.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of LREC: 11th International Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Anisia Katinskaia and Roman Yangarber. 2018. Digital cultural heritage and revitalization of endangered Finno-Ugric languages. In *Proceedings of the 3rd Conference on Digital Humanities in the Nordic Countries*, Helsinki, Finland.

Leonid Kharitonov. 1987. *Самоучитель якутского языка (Yakut language tutorial).* Yakutsk Publishing.

Evdokia Korkina. 1970. *Наклонения глагола в якутском языке.* Moscow, Nauka.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production.* Ph.D. thesis, Helsinki, Finland.

Nyurgun Leontiev. 2015. The newspaper corpus of the Yakut language. *TurkLang-2015*, page 233.

Krister Lindén, Erik Axelson, Sam Hardwick, Tommi A Pirinen, and Miikka Silfverberg. 2011. HFST—framework for compiling and applying morphologies. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 67–85. Springer.

Petr Sleptsov. 2018. *Саха тылын быһаарыылаах улахан тылдьыта: Большой толковый словарь якутского языка. (Large explanatory dictionary of the Yakut language: in 15 volumes).* Novosibirsk, Nauka.

Edhem Tenishev. 1997. *Языки мира: Тюркские языки (Languages of the World: Turkic languages)*, volume 2. Moscow: Indrik.

Elizaveta Ubryatova. 1982. *Грамматика современного якутского литературного языка (Grammar of the modern Yakut literary language: Phonetics and morphology).* Moscow, Nauka.

Rimma Vasilieva, M Degtyareva, N Ivanova, and L Semenova. 2013. *Современная этноязыковая ситуация в Республике Саха (Якутия): социопсихолингвистический аспект (Modern ethno-linguistic situation in the Republic of Sakha (Yakutia): sociopsycholinguistic aspect).* Novosibirsk, Nauka.

Roman Yangarber. 2018. Support for endangered and low-resource languages via e-learning, translation and crowd-sourcing. In *FEL XXII: Proceedings of the 22nd Annual Conference of the Foundation for Endangered Languages*, pages 90–97. London: FEL & EL Publishing.

Viktoria Yartseva. 1990. *Лингвистический энциклопедический словарь (Linguistic encyclopedic dictionary).* Moscow, Sovetskaya entsiklopediya.