Working Towards Digital Documentation of Uralic Languages With Open-Source Tools and Modern NLP Methods

Mika Hämäläinen[1], Jack Rueter [2], Khalid Alnajjar[3] and Niko Partanen[2]
[1] Metropolia University of Applied Sciences
[2] University of Helsinki
[3] Rootroo Ltd
firstname.lastname@metropolia.fi/helsinki.fi/rootroo.com

Abstract

We present our work towards building an infrastructure for documenting endangered languages with the focus on Uralic languages in particular. Our infrastructure consists of tools to write dictionaries so that entries are structured in XML format. These dictionaries are the foundation for rule-based NLP tools such as FSTs. We also work actively towards enhancing these dictionaries and tools by using the latest state-of-the-art neural models by generating training data through rules and lexica.

1 Introduction

Most of the languages spoken in the world are in danger of extinction. Their documentation and revitalization are of a highest cultural value, for which they have received plenty of academic attention in various disciplines such as anthropology, typology, lexicography and computational linguistics. Needless to say, the resources produced in each individual research project are not always published openly let alone made available to the community of native speakers.

The goal of our paper is to describe our open infrastructure for documenting minority languages. We present our experiences with the following Uralic languages: Skolt Sami (sms), Erzya (myv), Moksha (mdf), Komi-Zyrian (kpv) and Komi-Permyak (koi). As they belong to the Uralic branch, they are languages that exhibit a complex morphology, which makes their computational processing a challenge for modern machine learning methods that would require a lot of data to cover this complexity. The quantity and quality of data is usually an issue when we deal with endangered languages (Hämäläinen, 2021). Carrying out linguistic documentation in a structured machine readable format, however, makes it possible to create the resources needed for building NLP tools simultaneously with linguistic documentation.

We are about to start working with the Apurinã (apu) language, which allows us to reflect upon our Uralic context from a broader perspective, and increases the relevance of our work in a Latin American context. Thus, we describe how our infrastructure can work in non-Uralic contexts.

Linguistic documentation is a field of academic study that has developed considerably in recent decades. Its purpose is to provide a complete record of the linguistic practices characteristic to a given speech community (Himmelmann, 1998). The goal of language documentation is to describe the language of a community of speakers as fully as possible both for future generations and for language revitalization. The result of this work is typically manifested as a linguistic corpus or other type of material collection, which later on can be studied and analyzed in various ways. The question whether the collected materials actually describe the language use of a speech community is debatable, and this goal can never be fully achieved because a corpus can never describe a language in full. Nonetheless, linguistically collected materials may be the only resources available for a small language.

Whether and how language documentation materials should be made accessible and distributed, has been a matter of debate. We believe it is important to understand that this is also a matter of granularity, and the question is not necessarily whether the materials are accessible, but rather which parties should be allowed what type of access. There are good reasons for keeping culturally sensitive materials available only to specific groups. At the same time, there are always materials in any language that are more neutral and such that the authors themselves may want to make accessible. Especially for written publications, it may always be possible to negotiate a publication with open licenses, which would also allow the reuse of the same materials in different open research purposes.

Open materials are particularly important when we develop tools for NLP, because this work can greatly benefit from resources that are openly accessible with a permissive license. In the following sections we will discuss examples of such work, including our contribution to Universal Dependency treebanks. It must be emphasized that the open technology developed on an open infrastructure can also be used to process materials that are available only to a particular researcher or individual members of a community. Therefore, open infrastructure benefits both open and closed environments, whereas a closed infrastructure only benefits a big commercial player.

2 Related work

There are several individual projects in different parts of the world that work with online dictionaries for endangered languages. Many projects, however, focus on one language only and work without knowing about other ongoing projects for other endangered languages. This has led to a situation where researchers solve the same type of problems individually for their language of interest reinventing the wheel over and over again. There are plenty of online dictionaries and language learning tools that have been developed from scratch for one particular language.

Work with endangered languages in North America has shown the importance of language learning tools for second language learners. Lack of familiarity with lexicographical tradition can easily be a deciding factor in a beginner's learning experience. A learner of a new language cannot be expected to know exactly where an entry is located in a dictionary, nor can the learner be expected automatically to know the normative spelling. When the user of a language lacks a proper keyboard layout or knowledge of the correct orthography, the strategies of orthographic relaxation can be implemented in mobile and online dictionaries. Morphological processing and spelling relaxation are used to cater to beginners in Tsimshian and Salishan languages in the use of dictionaries and NLP tools (Littell et al., 2017).

On an entirely separate front, work has also been done to provide the Yupik community of St. Lawrence Island unimpeded access to language materials online. This has been possible using a morphologically aware dictionary. In the system, a strategy of multiple input methods has been introduced that caters to different writing systems (Hunt et al., 2019). The work here is tailor-made, and it maintains a strong link between the language and its community. The endangered language is seen as a low-resourced language in this context.

The problem is that *low-resourced language* is a term that is used for almost any language with less Internet presence than English. languages like Hindi (Irvine and Callison-Burch, 2014), Arabic (Chen et al., 2018) or Persian (Ahmadnia et al., 2017) are often considered low-resourced languages in the world of NLP, even though they have millions of speakers. In the work of Nasution et al. (2018), the ethnic Indonesian languages are relatively small compared to the superstrate language that surrounds them. The approach consists of working simultaneously with a group of closely related languages in a multilingual, language-independent infrastructure. The authors analyze the use of bilingual dictionary entries and explain the difficulty of selecting the appropriate bilingual dictionaries to begin documentation.

One of the largest infrastructures for minority language documentation from the point of view of computational linguistics is that of Giella (Moshagen et al., 2014). Their infrastructure is based on two main components: FST transducers (finite state transducers) and XML dictionaries. Transducers are a way of documenting the morphology of a language computationally. That is to say, they are collections of rules about how the morphological system of a language works. These rules can be used directly for automatic text analysis and lemma conjugation in its morphological variants.

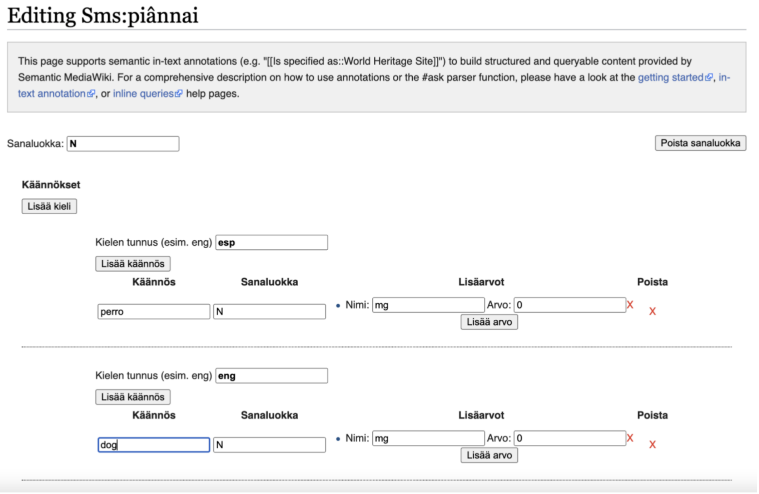
Transducers and XML dictionaries are used for spelling correction in Word[1], text prediction on Android and iOS keyboards[2], interactive systems to learn languages (Bontogon et al., 2018) and online dictionaries (Rueter and Hämäläinen, 2017). Our infrastructure is based on Giella, which allows us to synchronize data between the two infrastructures. This means that advances in linguistic documentation in our infrastructure can be used directly in the tools produced in Giella.

3 Our infrastructure

Using Giella requires a relatively high proficiency in programming to be able to write dictionaries and morphological rules for FSTs, and at the same time,

[1] http://divvun.no/korrektur/korrektur.html

[2] http://divvun.no/keyboards/mobileindex.html



Editing Sms:piânnai

This page supports semantic in-text annotations (e.g. "[[Is specified as::World Heritage Site]]") to build structured and queryable content provided by Semantic MediaWiki. For a comprehensive description on how to use annotations or the #ask parser function, please have a look at the getting started, in-text annotation, or inline queries help pages.

Sanaluokka: N

Käännökset

Kielen tunnus (esim. eng) esp

Käännös	Sanaluokka	Lisäarvot	Poista
perro	N	• Nimi: mg Arvo: 0	x

Kielen tunnus (esim. eng) eng

Käännös	Sanaluokka	Lisäarvot	Poista
dog	N	• Nimi: mg Arvo: 0	x

Figure 1: The form in Akusanat to edit the entry piânnai (dog) in Skolt Sami

it requires a good amount of knowledge in the language that is being documented. The infrastructure can be too complicated even for those who have studied computer science, and therefore it is not accessible to a community outside of those who collaborate directly with Giella. For this reason, our infrastructure has several interfaces for different types of users; for users who do not have sufficient knowledge to write XML or program transducers and for developers who want to use the tools without knowing how to compile them right from the beginning with the *make* command.

3.1 Online dictionaries

A very important step in the documentation of a minority language is the lexicographical work. This results in a dictionary that can be useful for both native speakers as for those who want to learn the language. We store dictionaries in a highly structured XML format. That means that all kinds of metadata are in their respective fields rather than being stored in various parts of a lexicographical entry in an unstructured format. This is important as we do not only want to create dictionaries for human use, but we also want them to be machine readable.

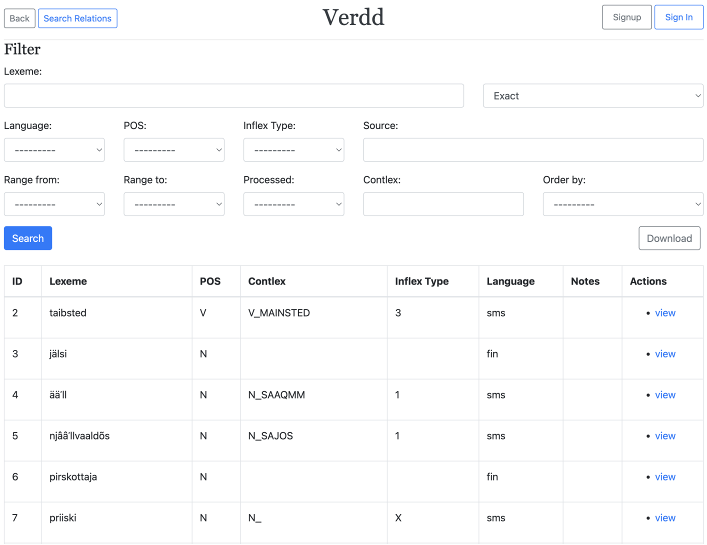
Our Akusanat system[3] (Hämäläinen and Rueter, 2018, 2019) is based on MediaWiki and allows you to view the content of XML dictionaries for all types of users. MediaWiki data is synchronized with XML files using the Git version control. This means that if someone modifies a lexicographical entry in Akusanat, these changes will result in a change to the XML dictionary stored in GitHub. If someone changes the XML dictionaries directly, Akusanat will download the new changes from GitHub and update its database automatically. This is done so that advanced users are able to edit the XML files directly with their favorite tool and less advanced users can make changes online with a graphical user interface. Akusanat does not let users modify the Wiki syntax directly, instead it displays a form that ensures changes remain structured and compatible with XML Figure 1.

For searching, we use morphological FST transducers to process the user input. This means that the user can search for a word in any of its morphological inflections, since the FST can lemmatize words automatically. It is also possible to search by typing in misspelled words. The transducers contain information about the most common spelling errors in each language, which allows us to resolve the lemma, although the word has not been spelled according to the spelling standard. This is important in the case of languages with which we work, since spelling rules are not as well-established as in the case of majority languages.

Figure 3 shows the interface for looking up words in the dictionaries. In the example, the search term is the Skolt Sami word *sooǥǥ*, which is the genitive of the word *sokk*, which means family. Our system lemmatizes the search term automatically with the Skolt Sami FST, and displays the input for the *sokk* lemma to the user.

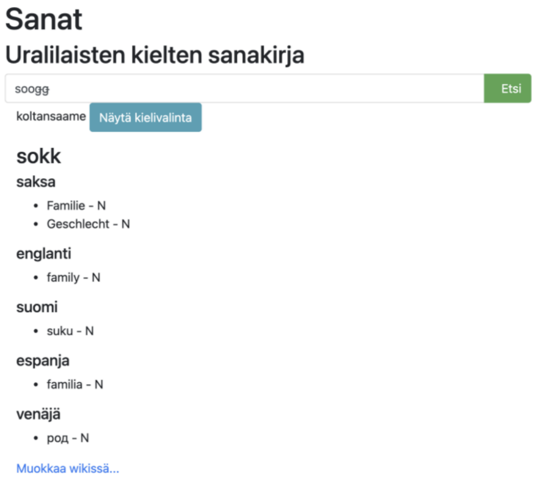
The idea of using MediaWiki, and especially Semantic MediaWiki, to create dictionaries, is not new, since there are already several projects that use the technology as their base (Muljadi et al., 2006; Bon and Nowak, 2013; Dueñas and Gómez,

[3] https://akusanat.com.



ID	Lexeme	POS	Contlex	Inflex Type	Language	Notes	Actions
2	taibsted	V	V_MAINSTED	3	sms		• view
3	jälsi	N			fin		• view
4	ää'll	N	N_SAAQMM	1	sms		• view
5	njââ'llvaaldõs	N	N_SAJOS	1	sms		• view
6	pirskottaja	N			fin		• view
7	priiski	N	N_	X	sms		• view

Figure 2: The interface for searching and filtering lexical entries in Ve'rdd



sokk

saksa

- Familie - N
- Geschlecht - N

englanti

- family - N

suomi

- suku - N

espanja

- familia - N

venäjä

- род - N

Figure 3: The interface for searching in Akusanat

2015). Without a doubt, MediaWiki has its advantages, in practice we have had to program our own MediaWiki extensions to add the necessary functionality; the form to edit, MediaWiki-XML synchronization, search with transducers etc. The problem that we have experienced many times is that the inner workings of MediaWiki change too often. This means that if we want to keep our MediaWiki instance up to date with the latest security updates, we have to make a lot of changes to our source code to keep our extensions working with the new version of MediaWiki. Even so, we continue to use and develop Akusanat[4] for the time being, as it offers a simple environment for users. In the next section, we describe the other system that we are developing. The new system may replace Akusanat in the future.

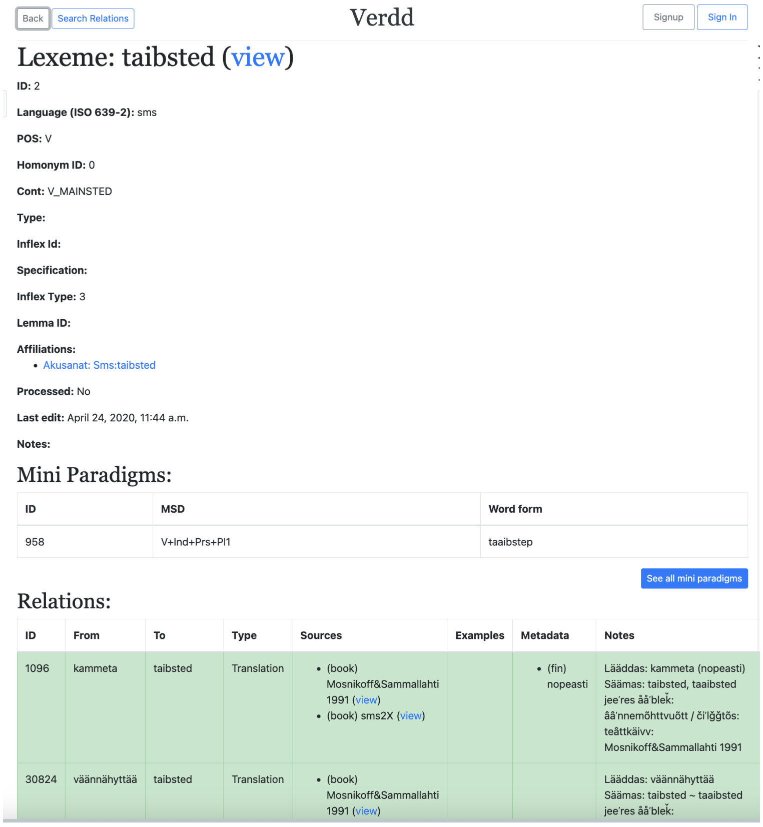
3.2 Editorial work

In this section, we describe the Ve'rdd[5] system (Alnajjar et al., 2019, 2020). The system works with the same XML dictionaries as Akusanat and can be used online in a similar way. The difference is in the intended use of the system. Ve'rdd is not a system to visualize lexicographical entries for an end user, but a system created specifically for writing both digital and printed dictionaries. During the process of developing the system, we have collaborated with a group of professional lexicographers who work with printed dictionaries.

In the context of the languages we work with, lexicographical documentation does not start from scratch, as both the Sami languages spoken in the Nordic countries and the Permian and Mordvinic languages spoken in Russia have received much attention in terms of their digital documentation during the last century. For example, there is a dictionary of the Skolt Sami language Sammallahti and Mosnikoff (1991), and there are several studies on the Mordvinic (Aasmäe et al., 2016; Grünthal, 2016) and Permian languages (Hamari, 2011; Klumpp, 2016). If there are existing dictionaries in digital form, they exist in an unstructured format such as a Word, CSV, or PDF file produced with an OCR system. For this reason, Ve'rdd includes functionality for import lexicographic data from unstructured formats. We have paid a lot of attention

[4] Code available https://github.com/mikahama/akusanat

[5] https://akusanat.com/verdd/



Lexeme: taibsted (view)

ID: 2

Language (ISO 639-2): sms

POS: V

Homonym ID: 0

Cont: V_MAINSTED

Type:

Inflex Id:

Specification:

Inflex Type: 3

Lemma ID:

Affiliations:

- Akusanat: Sms:taibsted

Processed: No

Last edit: April 24, 2020, 11:44 a.m.

Notes:

Mini Paradigms:

ID	MSD	Word form
958	V+Ind+Prs+Pl1	taaibstep

Relations:

ID	From	To	Type	Sources	Examples	Metadata	Notes
1096	kammeta	taibsted	Translation	• (book) Mosnikoff&Sammallahti 1991 (view) • (book) sms2X (view)		• (fin) nopeasti	Läddas: kammeta (nopeasti) Säämas: taibsted, taaibsted jee'res åå'blek: ââ'nnemõhttvuõtt / či'lǧǧtõs: teâttkäivv: Mosnikoff&Sammallahti 1991
30824	väännähyttää	taibsted	Translation	• (book) Mosnikoff&Sammallahti 1991 (view)			Läddas: väännähyttää Säämas: taibsted ~ taaibsted jee'res åå'blek:

Figure 4: The interface for editing lexical entries in Ve’rdd

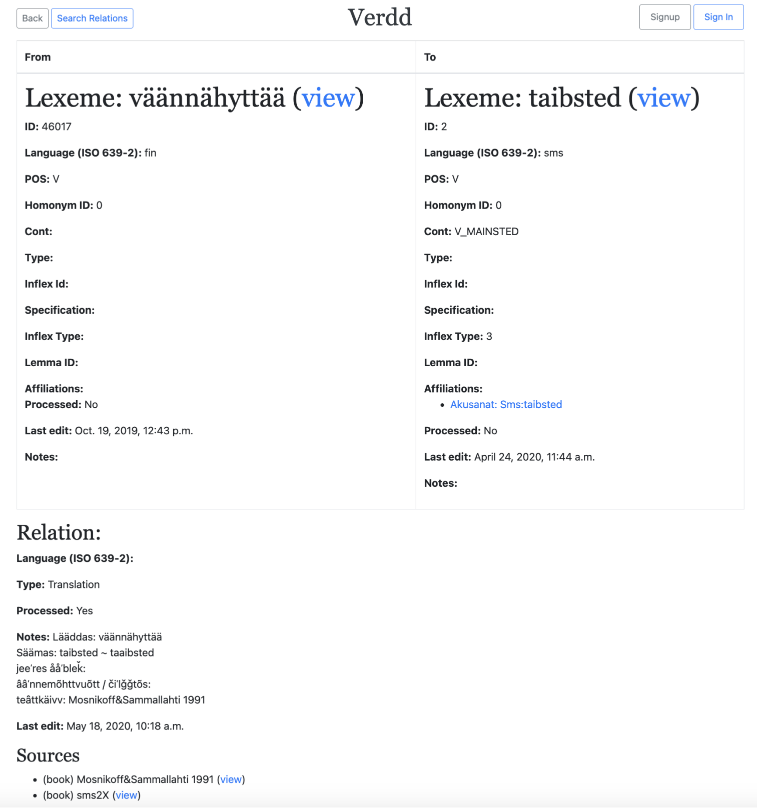
to the quality of the conversion, since, in the case of our languages, especially in the case of Skolt Sami, it is very frequent that the same character exists in many different Unicode characters. For example, ʹ (U+02B9 modifier letter prime) is a very common character in Skolt Sami, but because of the Finnish keyboard layout, it is often written as ' (U+0027 apostrophe) or ´ (U+00B4 acute accent). Ve’rdd is programmed to take into account the possible characters of the language and try to correct the incorrect characters automatically.

Figure 2 shows the interface for searching and filtering words in Ve’rdd. The interface is designed to support the workflow of a dictionary editor. For example, it is possible to display only raw inputs. This means entries that no one has verified after importing the data from an unstructured format. To facilitate the development of FST transducers it is also possible to sort and filter the words according to the continuation lexicon, which is the FST way of indicating how every word is supposed to be inflected.

Apart from just searching and filtering lexical entries, it is important to have the possibility to edit them. Figure 4 shows the interface for inspecting a dictionary entry. If a user is connected to their account, in addition to viewing, they can edit the information of a lexicographic entry. Ve’rdd is designed to be a tool for multilingual dictionaries, so one entry is connected to other entries in the system. In Figure 4, relationships can be seen as translation types that connect a word to its translations in other languages. It is also possible to define other types of relationships between lexica based on etymology. Relationships may also exist between words of the same language, for example, it is possible to indicate compound words and derivations with relations. Since the FST transducers contain derivative information, Ve’rdd automatically adds this type of relationship when importing a unstructured dictionary.

Ve’rdd can visualize the relationship between two words that are linked together with any kind of relationship. This can be used to verify that a word in a given language is linked to the correct homonym in another language (Figure 5). It is also possible to edit the type of relationship or delete any unnecessary relationships.

Ve’rdd has a functionality that allows the user to export any dictionary in different formats. The most important for us are the Giella XML, which can be used to generate FST transducers, and Latex



Back | Search Relations | Verdd | Signup | Sign In

From	To
Lexeme: väännähyttää (view)	Lexeme: taibsted (view)
ID: 46017	**ID:** 2
Language (ISO 639-2): fin	**Language (ISO 639-2):** sms
POS: V	**POS:** V
Homonym ID: 0	**Homonym ID:** 0
Cont:	**Cont:** V_MAINSTED
Type:	**Type:**
Inflex Id:	**Inflex Id:**
Specification:	**Specification:**
Inflex Type:	**Inflex Type:** 3
Lemma ID:	**Lemma ID:**
Affiliations:	**Affiliations:** • Akusanat: Sms:taibsted
Processed: No	**Processed:** No
Last edit: Oct. 19, 2019, 12:43 p.m.	**Last edit:** April 24, 2020, 11:44 a.m.
Notes:	**Notes:**

Relation:

Language (ISO 639-2):

Type: Translation

Processed: Yes

Notes: Läddas: väännähyttää
Säämas: taibsted ~ taaibsted
jee′res åå′blek̆:
ââ′nnemõhttvuõtt / či′lǧǧtõs:
teâttkäivv: Mosnikoff&Sammallahti 1991

Last edit: May 18, 2020, 10:18 a.m.

Sources

- (book) Mosnikoff&Sammallahti 1991 (view)
- (book) sms2X (view)

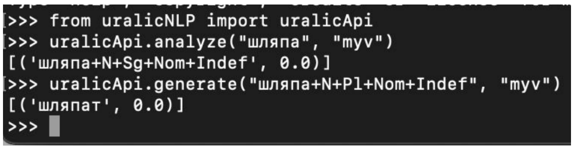
Figure 5: The interface for comparing two related entries in Ve'rdd

code. The Latex code makes it is possible to generate a ready-to-print PDF version of the dictionary. The Latex format makes it possible to change the style of the dictionary without changing the content. If there are changes in Ve'rdd, it is possible to update the content of the dictionary without changing the style defined in Latex. This functionality has been an important design principle for us since the work done in Ve'rdd should not only be used in digital dictionaries but also in printed dictionaries.

3.2.1 NLP resources

Our dictionary editing systems are directly useful in the development of FST transducers since we can export the lexicon in the format needed for HFST (Lindén et al., 2013). HFST is the tool we use to create the transducers. We have transducers for the Skolt Sami (Rueter and Hämäläinen, 2020), Erzya and Moksha (Rueter et al., 2020a) and Komi languages. The transducers can be used to lemmatize words, analyze their morphology or generate inflected forms. These transducers are difficult to compile for people who do not work with the transducers often. For this reason, we compile all transducers every night and we distribute them through our website[6]. We not only compile our transducers but all transducers for all languages in the Giella infrastructure.

The transducers are difficult to use as such, and for this reason, we have developed a Python library called UralicNLP (Hämäläinen, 2019) and a Python implementation of HFST called PyHFST (Hämäläinen and Alnajjar, 2023). With the libraries, compiled dictionaries and translators can be downloaded and used directly in Python. Fig 6 shows how to use our transducers from Python. In the second line of code, the word шляпа (hat) is analyzed in erzya (myv). The result indicates that the word is an indefinite (+Indef) noun (+N) in the nominative (+Nom) singular (+Sg). In the fourth line we generate the conjugated form of the same word in the plural (+Pl). The result is the plural word шляпат.

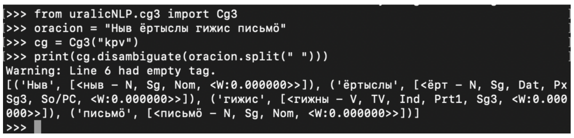


```
>>> from uralicNLP import uralicApi
>>> uralicApi.analyze("шляпа", "myv")
[('шляпа+N+Sg+Nom+Indef', 0.0)]
>>> uralicApi.generate("шляпа+N+Pl+Nom+Indef", "myv")
[('шляпат', 0.0)]
>>>
```

Figure 6: An example of using UralicNLP

[6] https://models.uralicnlp.com/nightly/

FST transducers produce all possible interpretations for a word from. In the case of the Uralic languages, there is plenty of homonymy in morphological inflections. This means that, if we use the transducers on regular text, we cannot accurately lemmatize the words in their context since the transducers produce all possible lemmas, For this reason, we use constraint grammar disambiguators (Karlsson et al., 2011) based on a tool called VISL CG-3 (Bick and Didriksen, 2015). The grammar rules of constraint grammars remove morphological readings that are not possible in a given sentence, and result in a sentence that is morphologically disambiguated with one lemma per word as opposed to all the possible lemmas.



```
>>> from uralicNLP.cg3 import Cg3
>>> oracion = "Ныв ёртыслы гижис письмӧ"
>>> cg = Cg3("kpv")
>>> print(cg.disambiguate(oracion.split(" ")))
Warning: Line 6 had empty tag.
[('Ныв', [<ныв - N, Sg, Nom, <W:0.000000>>]), ('ёртыслы', [<ёрт - N, Sg, Dat, Px
Sg3, So/PC, <W:0.000000>>]), ('гижис', [<гижны - V, TV, Ind, Prt1, Sg3, <W:0.000
000>>]), ('письмӧ', [<письмӧ - N, Sg, Nom, <W:0.000000>>])]
>>> 
```

Figure 7: An example of the use of the Komi Zyrian disambiguator

In Figure 7, we can see how the CG disambiguators can be used on UralicNLP. The third line initializes the disambiguation object for the Komi-Zyrian (kpv) and in the fourth line the disambiguation method of the object is called with a sentence. The result contains the word forms of the sentence, their lemmatization and morphology for each word of the sentence.

Apart from structured dictionaries and rule-based tools, we have treebanks of the universal dependencies for the Skolt Saami, Moksha, Erzya (Rueter and Tyers, 2018), Komi-Zyrian (Partanen et al., 2018) and Komi-Permyak (Rueter et al., 2020b). These treebanks contain syntactic annotations with the tags Morphological characteristics of universal dependencies. With the latest treebanks, we have also added the morphological labels produced by the transducers to facilitate the use of the two resources together

4 Incorporating modern NLP methods

As we have described thus far, a great part of our work relies on the old rule-based tradition of NLP. When we deal with endangered languages, rules are the primary starting point. One cannot simply train a neural network if there isn't enough training data. However, we do not want to reject neural models instantly as something that simply will not work for small languages. Neural models can work and they can be extremely beneficial. Throughout our research, we have aimed at combining rule-based models with neural models to facilitate our work on endangered languages.

Digital documentation has allowed us to use the latest methods in the world of NLP to automatically increase the data we have in the dictionaries. Because all of the lexicographic resources we have are multilingual, the first step we have taken with NLP technology has been the prediction of translations (Hämäläinen et al., 2018). The idea was as follows: if the Skolt Sami dictionary contains Finnish translations, German and English, and the Erzya dictionary contains translations into Finnish, English, Russian, and French, then, with this information, it should be possible to automatically deduce translations from Skolt Sami into Russian and French and from Erzya into German given the existence of two common languages: Finnish and English. With a probabilistic model we have increased the number of translations in Skolt Sami, Erzya, Moksha and Komi-Zyrian dictionaries.

We have elaborated on this idea later on by using graph based approaches and neural models (Alnajjar et al., 2021, 2022). These have not been isolated attempts, but the graph based methods have been incorporated into Ve'rdd as well. The predictions have been manually checked and this way we have been able to augment our dictionaries semi-automatically. The Livonian institute has embraced this technology in bootstrapping a Livonian-English dictionary.

As neural networks require a large amount of data to be trained, it is common to believe that their use is not possible in the case of endangered languages. We have taken the perspective that we can generate the amount of data needed for a neural network with our morphological tools. Using the treebanks and the transducers, we have generated data to train a neural network to perform disambiguation instead of using the constraint grammar for Erzya and Komi-Zyrian (Ens et al., 2019). The idea was to generate all possible analyzes for the words in the treebanks and train the neural network to disambiguate the analyses with the treebank analysis. Later on, we further developed this method in the context of Sami languages (Hämäläinen and Wiechetek, 2020).

We have also been able to use the neural networks to increase etymological relationships in the Skolt Sami dictionary (Hämäläinen and Reuter, 2019). The method was based on a character level LSTM model that was enhanced with synthetic data generated with a character-level statistical machine translation tool. We used this method to produce a set of candidate cognates that we manually checked and incorporated into our digital dictionaries. This method relies on external data from the Institute for Languages in Finland, which makes it difficult for us to include it in Ve'rdd.

Rule-based FSTs are great because they are usually very accurate, however, they do not have a great lexical coverage. Analyzing an online text with the FSTs will usually mean a ton of words that are not recognized at all. For this reason, we used the FSTs to generate training data for neural models (Hämäläinen et al., 2021). We used this data to train character-level neural machine translation models to analyze, generate and lemmatize word forms. The key idea is to use the exact same morphological tags so that the neural models and the FSTs can be used interchangeably. These neural models have been made available through UralicNLP as a fallback mechanism. If an FST fails to analyze a word form, the neural model will be used automatically if neural fallback is turned on.

Recently, we have also moved our interest towards other aspects of NLP than just lexicon and morphology. We have done work on automatically translating and aligning word embeddings for endangered Uralic languages (Alnajjar, 2021) and using them successfully in downstream tasks such as sentiment analysis (Alnajjar et al., 2023).

5 Discussion and Conclusions

We hope that our work can be useful for others as well. We have put a lot of attention in open-sourcing our tools and resources so that nobody needs to start building language documentation tools entirely from scratch. We have also paved a road towards using state-of-the-art neural models in the context of truly endangered languages with extremely limited resources. This is challenging and requires ingenuity. We are not interested in committing to the dichotomy of researches who defend rule-based tools as the only viable option for endangered languages nor to the researchers who frown upon rules and rely solely on the Transformer architecture. The best solutions, we believe, are found by combining both worlds.

Our tools are compatible with the Giella infrastructure. This has made it possible to use our dictionaries and translators directly on their online platform to learn languages (Antonsen and Argese, 2018), on Android and iPhone keyboards and spell checking for Word and OpenOffice developed by Divvun[7] at Giella. Flexible and interoperable design makes it also possible to integrate different lexical resources into our infrastructure once those are digitized or otherwise become available.

Digital documentation clearly has its benefits, since we can carry out machine learning with structured dictionaries and FST transducers. For this reason, a project conducted at the University of Oulu[8] the goal of which was to author the new dictionary Skolt Finnish-Sami has chosen to use Ve'rdd to create the digital and printed dictionary. We have worked together with project employees to increase the functionality of our system. Ve'rdd has made the simultaneous work of editors possible who, without Ve'rdd, would have used Excel and Word. This would have meant a lost chance of producing a structured dictionary for the interest of NLP and a printed dictionary at the same time.

We have started to explore non-Uralic languages by building a UD treebak for Apurinã (Rueter et al., 2021). Furthermore, we have built an initial FST for Lushootseed (lut) (Rueter et al., 2023) and extended it with an LSTM model. These are our initial steps towards non-Uralic languages.

References

Niina Aasmäe, Karl Pajusalu, and NADEŽDA KABAJEVA. 2016. Gemination in the mordvin languages. *Linguistica Uralica*, 52(2).

Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian-spanish low-resource statistical machine translation through english as pivot language. In *RANLP*, pages 24–30.

Khalid Alnajjar. 2021. When word embeddings become endangered. In *Multilingual Facilitation*, pages 275–288. University of Helsinki.

Khalid Alnajjar, Mika Hämäläinen, Niko Partanen, and Jack Rueter. 2019. The open dictionary infrastructure for uralic languages. Электронная Письменность Народов Российской Федерации.

Khalid Alnajjar, Mika Hämäläinen, Niko Tapio Partanen, and Jack Rueter. 2022. Using graph-based methods to augment online dictionaries of endangered languages. In *Proceedings of the Fifth Workshop on*

[7] http://divvun.no/

[8] https://www.sttinfo.fi/tiedote/tekoaly-apuna-koltansaamen-ja-pohjoissaamen-digitaalisten-sanakirjojen-toimitustyossa?publisherId=57858920&releaseId=69886820

the Use of Computational Methods in the Study of Endangered Languages, pages 139–148.

Khalid Alnajjar, Mika Hämäläinen, and Jack Rueter. 2023. Sentiment analysis using aligned word embeddings for uralic languages. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 19–24.

Khalid Alnajjar, Mika Hämäläinen, Jack Rueter, and Niko Partanen. 2020. Ve'rdd. narrowing the gap between paper dictionaries, low-resource nlp and community involvement. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 1–6.

Khalid Alnajjar, Jack Rueter, Niko Partanen, and Mika Hämäläinen. 2021. Enhancing the erzya-moksha dictionary automatically with link prediction. *Folia Uralica Debreceniensia*.

Lene Antonsen and Chiara Argese. 2018. Using authentic texts for grammar exercises for a minority language. In *Proceedings of the 7th workshop on NLP for Computer Assisted Language Learning*, pages 1–9.

Eckhard Bick and Tino Didriksen. 2015. Cg-3—beyond classical constraint grammar. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 31–39.

Bruno Bon and Krzysztof Nowak. 2013. Wikilexicographica. linking medieval latin dictionaries with semantic mediawiki. In *eLex 2013*, pages 407–420. Trojina, Institute for Applied Slovene Studies (Ljubljana, Slovenia); Eesti . . . .

Megan Bontogon, Antti Arppe, Lene Antonsen, Dorothy Thunder, and Jordan Lachler. 2018. Intelligent computer assisted language learning (icall) for nêhiyawêwin: an in-depth user-experience evaluation. *Canadian Modern Language Review*, 74(3):337–362.

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

George Dueñas and Diego Gómez. 2015. A bilingual dictionary with semantic mediawiki: The language saliba's case. *The 4th International Conference on Language Documentation and Conservation (ICLDC)*.

Jeff Ens, Mika Hämäläinen, Jack Rueter, and Philippe Pasquier. 2019. Morphosyntactic disambiguation in an endangered language setting. In *22nd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Linköping University Electronic Press.

Riho Grünthal. 2016. Transitivity in erzya: Second language speakers in a grammatical focus. *Mordvin languages in the field*.

Mika Hämäläinen. 2019. Uralicnlp: An nlp library for uralic languages. *Journal of open source software*.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In *Multilingual Facilitation*. University of Helsinki.

Mika Hämäläinen, Niko Partanen, Jack Rueter, and Khalid Alnajjar. 2021. Neural morphology dataset and models for multiple languages, from the large to the endangered. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 166–177.

Mika Hämäläinen and Jack Rueter. 2019. An open online dictionary for endangered uralic languages. *Electronic lexicography in the 21st century*.

Mika Hämäläinen and Jack Michael Rueter. 2018. Advances in synchronized xml-mediawiki dictionary development in the context of endangered uralic languages. In *Proceedings of the XVIII EURALEX International Congress*. Ljubljana University Press.

Mika Hämäläinen, Liisa Lotta Tarvainen, and Jack Rueter. 2018. Combining concepts and their translations from structured dictionaries of uralic minority languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Mika Hämäläinen and Linda Wiechetek. 2020. Morphological disambiguation of south sámi with fsts and neural networks. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 36–40.

Arja Hamari. 2011. The abessive in the permic languages. *Suomalais-Ugrilaisen Seuran Aikakauskirja*, 2011(93):37–84.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1).

Benjamin Hunt, Emily Chen, Sylvia LR Schreiner, and Lane Schwartz. 2019. Community lexical access for an endangered polysynthetic language: An electronic dictionary for st. lawrence island yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 122–126.

Mika Hämäläinen and Khalid Alnajjar. 2023. Pyhfst: A pure python implementation of hfst. *Zenodo*. 10.5281/zenodo.7791470.

Mika Hämäläinen and Jack Reuter. 2019. Finding sami cognates with a character-based nmt approach. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Ann Irvine and Chris Callison-Burch. 2014. Hallucinating phrase translations for low resource mt. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 160–170.

Fred Karlsson, Atro Voutilainen, Juha Heikkilae, and Arto Anttila. 2011. *Constraint Grammar: a language-independent system for parsing unrestricted text*, volume 4. Walter de Gruyter.

Gerson Klumpp. 2016. Semantic functions of complementizers in permic languages. *Complementizer Semantics in European Languages*, pages 529–586.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Juha Kuokkala, Jyrki Niemi, Tommi A Pirinen, and Miikka Silfverberg. 2013. Hfst—a system for creating nlp tools. In *Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings 3*, pages 53–71. Springer.

Patrick Littell, Aidan Pine, and Henry Davis. 2017. Waldayu and waldayu mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 141–150.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M Tyers. 2014. Open-source infrastructures for collaborative work on under-resourced languages. *Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, pages 71–77.

Hendry Muljadi, Hideaki Takeda, Shoko Kawamoto, Satoshi Kobayashi, and Asao Fujiyama. 2006. Towards a semantic wiki-based japanese biodictionary. In *SemWiki*.

Arbi Haza Nasution, Yohei Murakami, and Toru Ishida. 2018. Designing a collaborative process to create bilingual dictionaries of indonesian ethnic languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first komi-zyrian universal dependencies treebanks. In *Second Workshop on Universal Dependencies (UDW 2018), November 2018, Brussels, Belgium*, pages 126–132.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies treebank. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.

Jack Rueter and Mika Hämäläinen. 2017. Synchronized mediawiki based analyzer dictionary development. In *3rd International Workshop for Computational Linguistics of Uralic Languages (IWCLUL 2017)*. The Association for Computational Linguistics.

Jack Rueter and Mika Hämäläinen. 2020. Fst morphology for the endangered skolt sami language. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 250–257.

Jack Rueter, Mika Hämäläinen, and Khalid Alnajjar. 2023. Modelling the reduplicating Lushootseed morphology with an FST and LSTM. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 40–46, Toronto, Canada. Association for Computational Linguistics.

Jack Rueter, Mika Hämäläinen, and Niko Partanen. 2020a. Open-source morphology for endangered mordvinic languages. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*. The Association for Computational Linguistics.

Jack Rueter, Niko Partanen, and Larisa Ponomareva. 2020b. On the questions in developing computational infrastructure for komi-permyak. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 15–25.

Jack Rueter and Francis Tyers. 2018. Towards an open-source universal-dependency treebank for erzya. In *Proceedings of the Fourth International Workshop on Computational Linguistics of Uralic Languages*, pages 106–118.

Pekka Sammallahti and Jouni Mosnikoff. 1991. *Suomi-koltansaame sanakirja: Lää'dd-sää'm sää'nnke̮'rjj*. Girjegiisá.