An Online Tool Developed for Post-Editing the New Skolt Sami Dictionary

Mika Hämäläinen[1], Khalid Alnajjar[1], Jack Rueter[1], Miika Lehtinen[2] and Niko Partanen[1]

[1]University of Helsinki, Unioninkatu 40, 00100 Helsinki, Finland
[2] University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland
E-mail: firstname.lastname@helsinki.fi, firstname.lastname@oulu.fi

Abstract

In this paper, we present our free and open-source online dictionary editing system that has been developed for editing the new edition of the Finnish-Skolt Sami dictionary. We describe how the system can be used in post-editing a dictionary and how NLP methods have been incorporated as a part of the workflow. In practice, this means the use of FSTs (finite-state transducers) to enhance connections between lexemes and to generate inflection paradigms automatically. We also discuss our work in the wider context of lexicography of endangered languages. Our solutions are based on the open-source work conducted in the Giella infrastructure, which means that our system can be easily extended to other endangered languages as well. We have collaborated closely with Skolt Sami community lexicographers in order to build the system for their needs. As a result of this collaboration, the latest Finnish-Skolt Sami dictionary was edited and published using our system.

Keywords: Skolt Sami, online dictionary, NLP

1. Introduction

In this paper, we present an online system developed in close collaboration with linguists and native speakers during the Skolt Sami dictionary project (see Alnajjar et al. 2020). We recognise that when developing lexical resources for endangered languages we must take into account various user groups and their needs, and the resource that is created is often in a very important position for the entire language community. Large dictionaries in endangered languages often play an important role in the future language development and efforts at normalisation. This means that these projects entail lots of responsibility. Establishing a common ground with knowledgeable native speakers and pencil and paper linguists with regard to online editing can present quite a challenge. Native speakers, on the one hand, need to be given an understandable and intuitive system for interacting with the growing dictionary database. Experienced linguists, on the other, may at times require an outstretched hand of enlightenment, one that introduces them to direct work in a database without interceding paper prints for contemplation of all entries with a pencil and eraser. The developers, of course, must also be prepared to design print-out and download possibilities just in case the users have difficulties managing the computer-readable data. In these instances the exported versions should also be used primarily to read and use the dictionary, and the changes should be done in the actual database, if possible.

Only this way we can ensure that the lexical resources that are being created will definitely benefit different user groups, and take into account the multiple purposes these materials can be used for. We also acknowledge that there is a need for specialised lexicographic solutions in different situations, and that the work presented here on Skolt Sami is just one of the many possibilities. At the same time there are many important lessons to be learned from our Skolt Sami work, and these can be generalised in different scenarios.

The work with Skolt Sami was started using a tabular data format. Spreadsheet editing programs are readily available and many linguists as well as native speakers are familiar with them, so it is obvious many endangered language lexicons appear in such formats. For

this reason, our system has also been designed so that these can be processed. Converting different tabular files is often not trivial, and this was the case in this project too, with the original lexical resources for Skolt Sami presenting several challenges. The most prominent consisted of a malformed flat CSV file containing several character encoding issues. We built our online system so that it fixes such issues while importing the flat structure into its internal graph based representation. Similar issues are common for different materials on endangered languages, so our solutions generalise very well to this wider context. We use graphs as the internal structure for their advantages over trees (see Mechura 2016). Despite the popularity of spreadsheets, this structure is poorly suited to lexicographic work. There are relations between entries, hierarchical entries, and additional content such as example sentences that can serve as examples for multiple different headwords, in which case repeating them again and again is not desirable. Lexicographic data is by nature relatively complicated to model, but as we will describe, the approach to import tabular formats into our online tool seems to provide a very good starting point for the creation of such a more complex structure, partly through automatic conversions and deductions.

Even though Skolt Sami is severely endangered with its 300 native speakers (Moseley, 2010), thanks to previous projects on its digital revitalisation, the language has morphological analysers (Rueter & Hämäläinen, 2020) that our system can use when importing data. Our system will automatically add relation information such as derivations and compounds to lexemes with the help of the morphological tools. If the system were to be used for a language that does not have a morphological analyser, these relations would need to be created either manually or by using different heuristics. In any case, the resulting dictionary would not be as interlinked.

Our system has been in continuous use by linguists and trained native speakers, who have been editing the lexicographic material into a publishable form. We have introduced constant improvements to the system based on the feedback from our actual users. Some of the requested functionalities have been automatic morphological inflections for full inflectional paradigms for each entry with a feedback facility, the ability to have an overseer view where a super user can see the edit history of each entry and finalise/approve it, and the ability of showing lexicographic information from other sources, such as the Sami TermWiki[1].

The final product, a printed edition of the dictionary (Lehtinen et al., 2021) was recently published, and it was greatly facilitated by the fact that our system can output the desired lexicographic content in a LaTeX format for easy PDF conversion. Other output formats could be easily added, if needed by the community or researchers.

Currently, we are extending the use of our system to other endangered languages documented in the Giella infrastructure (Moshagen et al., 2014). Like Skolt Sami, these languages have morphological tools as well, which makes work with them analogous to what we have already developed for Skolt Sami.

2. Related Work

Developing dictionaries is essentially connected to language documentation and revitalisation activities in the contemporary world. With entirely undocumented languages

[1] https://satni.uit.no/termwiki/

the lexicon is built from scratch as part of the corpus building and elicitation process, whereas in many cases there are existing dictionaries and lexical resources that can be used. Common approaches are to extend existing resources, or to publish them again in a digital format. There is also extensive global variation in what kind of resources exist and what kind of challenges are connected to making them usable for the communities. We will describe some of the most relevant work next.

Especially with the work on endangered languages of North America there are many examples where unfamiliarity with the orthographic conventions of the language is an issue in language learning. Additionally, many orthographic norms are not entirely fixed, if they exist at all, which is a challenge for lexicographic work. It is also a problem for a new use of the lexical infrastructure, as the user cannot be expected to know how to find a specific entry in the dictionary. Both spell relax and morphological awareness are methods that have been used in Tsimsianic and Salish dictionaries, with the aid of language technology that has been developed for these languages (Littell et al., 2017).

Another example comes from work done with St. Lawrence Island Yupik, where the language materials have been made openly available for the community online. Different writing systems that have previously been used for this language have been taken into account as different input methods, also here with the aid of morphological modelling (Hunt et al., 2019). As similar situations with various writing systems is very common for endangered languages around the world, and there are various ways to handle this issue. Situations are also different, since in some contexts different writing systems are actively in use, whereas at times they represent different historical periods of orthography development. One approach that has been designed for some endangered languages of Russia is to develop separate transliteration conventions between different writing systems, to the extent that is possible (Bradley & Skribnik, 2021).

One challenge we also identify is that the concept of a low-resource language is often used in a very inexact manner, as discussed further by Hämäläinen (2021). Any language besides English can in some situation be called a low-resource language, which makes the category difficult to use, and the concept less practical. Still, there are important differences between languages and the existing resources for them. This governs the starting point for further work, which makes it important to be able to contextualise up to some degree. Building new lexical resources is an entirely different undertaking when other bi- or multilingual lexicons already exist, even though they would differ in various ways from a new resource currently planned. In a study by Nasution et al. (2018) existing bilingual dictionaries in individual languages were used to create new resources for different language pairs. Even in this case, some of the languages were significantly smaller than the majority languages, which were also included in the original dataset.

Our method relies heavily on an existing morphological analyser. Such tools are not available for all languages, but the number of languages with at least some degree of coverage is not small, even if we look into individual infrastructures such as GiellaLT[2], or a Python package that can access these and other analysers, described by Hämäläinen (2019). At the same integrating the development work of a morphological analyser into the whole language documentation work and dictionary creation is not unprecedented either. Pirinen 2019 has reported in detail his parallel work on Karelian treebanks,

[2] https://github.com/giellalt/

dictionaries and computational grammar. As similar approach where a morphological analyser supports language documentation work is reported in Gerstenberger et al. (2017), although this did not include a more specific discussion about lexicographic work, which is still connected to the creation of an analyser on at least the lemma level. Wilbur (2017) developed a workflow for Pite Sami where lexicographic data is stored in a database and connected to the morphological analysis, which provides a strong parallel to our work.

Lexonomy (Měchura et al., 2017) is a good all-purpose online tool for dictionary editing. However, it is not sufficient for our needs. The main reason is that our aim is to have the system built in such a fashion that it can be directly used with the existing tools for Uralic languages (XML dictionary conventions, FST morphology and so on) (see Pirinen & Tyers 2021). We also need to provide an interface for users who are not familiar with the technology, and even the mere fact of having the XML structure visible in an advanced view might startle them.

We must also emphasise that often the endangered languages with limited resources do not have a native speaker base who could participate in the lexicographical work. This also calls for very customised and specialised solutions in each situation. We see, however, that there are some general characteristics and demands upon which the specialised versions can be constructed, instead of designing everything from scratch.

3. Our Online Editor

In this section, we describe our online dictionary editor. It is fully open source[3] and based on technologies such as Django[4] and the MariaDB database[5]. One of the key design goals of the editor has been building it on top of Giella's (Moshagen et al., 2014) reusable components, this means that the system can input and output Giella formatted XML dictionaries and use the NLP tools provided in the infrastructure.

3.1 User Interface

Our online system is bundled with numerous features and commands to facilitate searching, editing and producing dictionaries. These features include, but are not limited to, importing and exporting dictionaries from Giella's XMLs and CSVs, merging and cleaning lexemes, searching and approving entries in the dictionary, and generating a printable dictionary in LaTeX. In this section, we show a glimpse of the user-interface.

Figure 1 displays the homepage of the system where users can perform simple and complex search queries to find lexemes and interesting patterns. Simple filtering involves matching lexemes that either contain, start or end with, or have an exact match with the input query, whereas complex filtering can be conducted with the help of regular expressions (e.g., matching lexemes following a given pattern such as starting with "v" and ending with "ed"). Further filtering, for instance based on the part-of-speech, language, the source of the lexeme and/or whether it has been checked by an expert in the language, can be applied to retrieve relevant lexemes promptly.

[3] A GitHub link will be provided in the camera ready version
[4] https://www.djangoproject.com/
[5] https://mariadb.org/

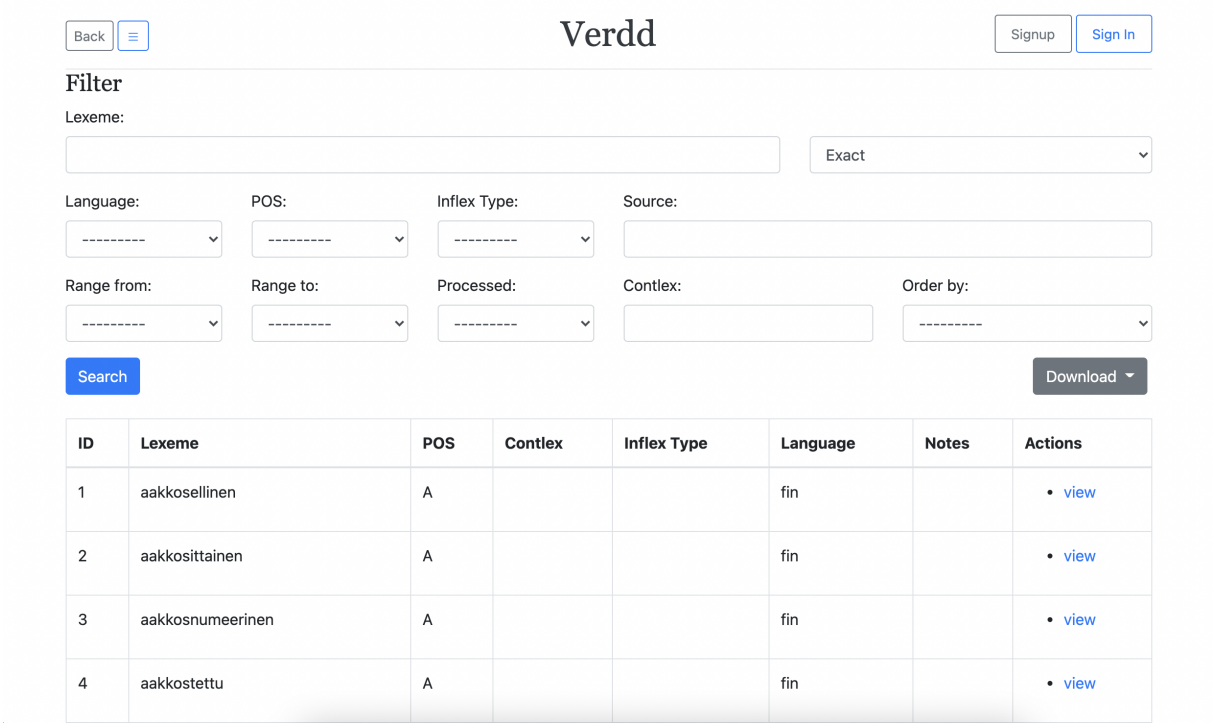


Figure 1: The user-interface for searching for lexemes in Ve′rdd.

When a user navigates to a given lexeme, all the information regarding the lexeme along with all relations to and from it are returned. An example of what is supplied to the user when visiting a lexeme is given in Figure 2. In this example, the lexeme is "ve′rdd". In addition to the core information of the lexeme (e.g., its language, POS and notes), our online system utilises FSTs (Finite-State Transducers) dedicated to the language to produce mini- and full- paradigms of the language. The user has the ability to override any automatically generated paradigms or even introduce new ones, which would serve as a feedback interface for improving the state of the FST. At the end of the lexeme page, all of its relations, e.g., derivations and translations, are shown, along with any examples and metadata which might be present for each relation.

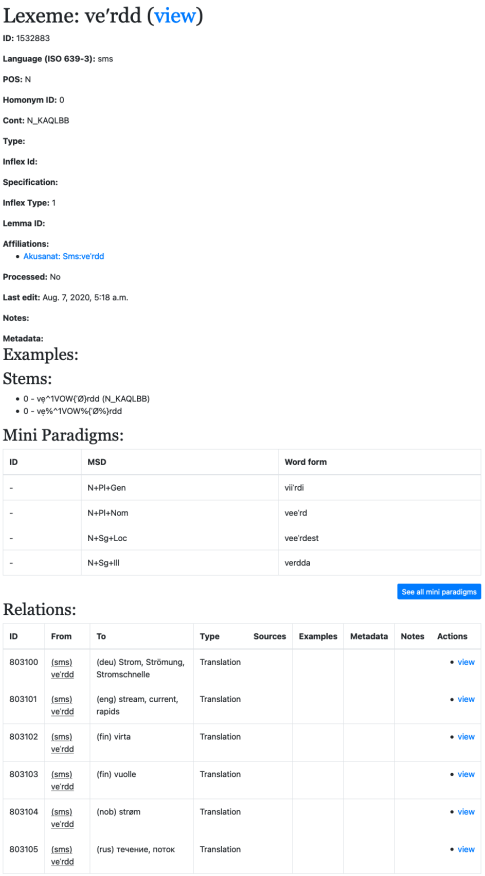
3.2 The data structure

In our system, the basic unit is a lexeme. A lexeme is just a word consisting of its lemma, part-of-speech and other metadata. If there are two words, the lemmas of which are homonyms, they will be two separate lexemes in the system with distinctive homonym IDs. *Sokk* is an example of such a case. It can be a word for *a family* or *a sock*, but it is inflected differently depending on which one of the homonyms is in question.

Lexemes are linked to each other with relations. These can be virtually anything, but in practice we have translation, derivation, compound and etymological relations. Relations can be uni- or bidirectional.

3.3 Importing and Exporting Data

Since the very beginning of the Skolt Sami dictionary project, it was evident that the system needed to support multiple different input formats. On the one hand, the original



Lexeme: ve′rdd (view)

ID: 1532883

Language (ISO 639-3): sms

POS: N

Homonym ID: 0

Cont: N_KAQLBB

Type:

Inflex Id:

Specification:

Inflex Type: 1

Lemma ID:

Affiliations:
- Akusanat: Sms:ve'rdd

Processed: No

Last edit: Aug. 7, 2020, 5:18 a.m.

Notes:

Metadata:

Examples:

Stems:
- 0 - ve^1VOW{Ø}rdd (N_KAQLBB)
- 0 - ve%^1VOW%{Ø%}rdd

Mini Paradigms:

ID	MSD	Word form
-	N+Pl+Gen	vii'rdi
-	N+Pl+Nom	vee'rd
-	N+Sg+Loc	vee'rdest
-	N+Sg+Ill	verdda

See all mini paradigms

Relations:

ID	From	To	Type	Sources	Examples	Metadata	Notes	Actions
803100	(sms) ve'rdd	(deu) Strom, Strömung, Stromschnelle	Translation					• view
803101	(sms) ve'rdd	(eng) stream, current, rapids	Translation					• view
803102	(sms) ve'rdd	(fin) virta	Translation					• view
803103	(sms) ve'rdd	(fin) vuolle	Translation					• view
803104	(sms) ve'rdd	(nob) strøm	Translation					• view
803105	(sms) ve'rdd	(rus) течение, поток	Translation					• view

Figure 2: Information displayed to the user when accessing a lexeme, “ve′rdd” in this case.

material of the first Finnish-Skolt Sami dictionary (Sammallahti & Mosnikoff, 1991), which was stored in a CSV format, needed to be imported, on the other hand, we needed to import the latest advances in the Giella XML-based Skolt Sami dictionary[6].

The first issue was the inconsistent characters that were used; the recent XML dictionaries only consisted of correct characters without an extended vocabulary, while the older CSV material had many different wrong encodings. For example, Skolt Sami uses the modifier letter prime character in its orthography in a word such as *ve′rdd* (stream), however words containing this character were often written with a single quote *ve'rdd* or as an accent *ve´rdd*. For this reason, we implemented a feature in our system that takes in a list of accepted characters in the language one is importing and shows an error if an unaccepted character is being imported. The system also takes in a conversion map that it uses to resolve erroneous characters automatically.

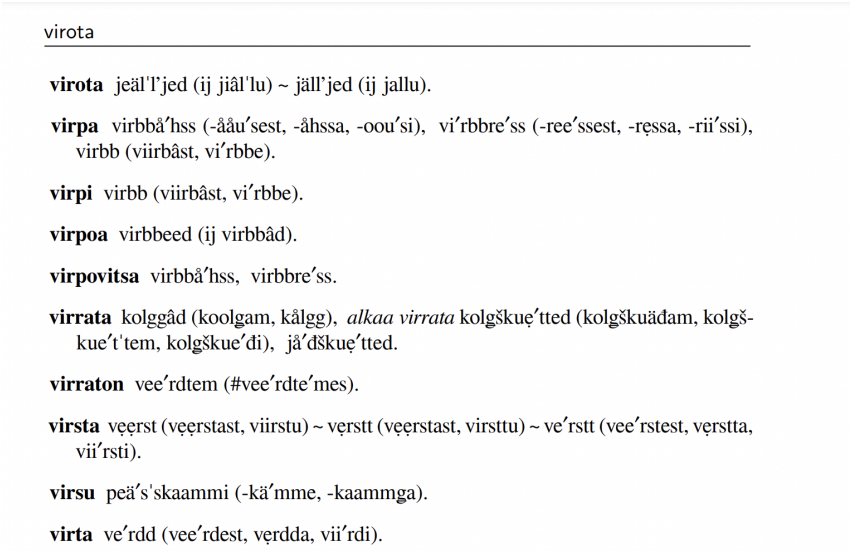
When the data was imported, we needed to support several output formats. First and foremost, Giella XML. This format is needed because several tools such as spell checkers and online language learning tools use dictionaries in this format. This means that this output format makes it possible for us to upload changes made in our system to the Giella infrastructure to benefit the higher level tools of the infrastructure.

Other output formats needed were CSV format as some lexicographers found it easier to work on that format as well, and most importantly LaTeX for producing the final

[6] https://gtsvn.uit.no/langtech/trunk/words/dicts/sms2X/

printable dictionary. The LaTeX output is generated with Django's template language[7], this means that customising the output dictionary does not require modifications to the program logic of our system, merely edits in the template file.

The interface allows downloading a printable dictionary edition in LaTeX format. Figure 3 shows part of a page of the printable dictionary that is automatically generated by our system. Our LaTeX template takes care of all the essential printed-dictionary formatting requirements, such as dividing the dictionary into alphabetised chapters, adding page headers containing guiding words and allowing single- or double- column dictionaries. The PDF output of the printable dictionary is searchable using any PDF reader, which permits distributing two versions of the dictionary: 1) an electronic version that is properly built and indexed, and 2) a physical dictionary.



virota

virota jeäl'l'jed (ij jiâl'lu) ~ jäll'jed (ij jallu).

virpa virbbå′hss (-åâu′sest, -åhssa, -oou′si), vi′rbbre′ss (-ree′ssest, -ręssa, -rii′ssi), virbb (viirbâst, vi′rbbe).

virpi virbb (viirbâst, vi′rbbe).

virpoa virbbeed (ij virbbâd).

virpovitsa virbbå′hss, virbbre′ss.

virrata kolggâd (koolgam, kålgg), *alkaa virrata* kolgškuę′tted (kolgškuäđam, kolgškue′t'tem, kolgškue′đi), jå′đškuę′tted.

virraton vee′rdtem (#vee′rdte′mes).

virsta vęęrst (vęęrstast, viirstu) ~ vęrstt (vęęrstast, virsttu) ~ ve′rstt (vee′rstest, vęrstta, vii′rsti).

virsu peä′s'skaammi (-kä′mme, -kaammga).

virta ve′rdd (vee′rdest, vęrdda, vii′rdi).

Figure 3: A snapshot of a page in the automatically produced printable dictionary.

3.4 Integration with NLP Tools

Our system uses FSTs (Finite-State Transducers) based on a tool called HFST by Lindén et al. (2013). These are useful as they produce morphological readings for word forms and they can be used to generate inflectional forms based on a lemma and morphological tags. We use these FSTs for two purposes: inflection and relations.

When a new word is input into the system, the first thing the system does is that it consults the FST and sees if this word is a derivational form of another word or if the new word is a compound formed of existing words. The system will then suggest to the person editing the dictionary that derivational and compound relations can be added automatically. All the editor needs to do is to either confirm or reject the automatically produced relations.

An important part of a dictionary of any morphologically rich language is the presence of certain inflectional forms in the lexicographic entry as, based on them, the user can

[7] https://docs.djangoproject.com/en/3.1/ref/templates/language/

know the full inflectional paradigm (see Hulden & Silfverberg 2021). We generate these inflectional forms automatically in our system for all input words. These can be inspected under the miniparadigm field. The dictionary editor can override these automatically generated inflectional forms by editing them. This also serves as feedback for the people editing the FSTs so that they can correct any mistakes in their output.

We are currently integrating our latest graph and deep learning based methods (Hämäläinen et al., 2021) into our system. We have been able to automatically predict new translations for the Giella dictionaries based on XML dictionaries in other languages and Wikitionaries in large languages. In short, for a lemma that has translations into at least two other languages, our method can predict more suitable synonymous translations for the two languages and translation candidates with the same meaning in other languages, with the idea that the more languages our system covers, the more nuanced its understanding of polysemy becomes.

4. Discussion and Future Directions

In the future, the dictionary editing platform has to be tested with different languages and editorial teams. This is necessary so that we understand what kinds of workflows serve different communities best, and which of the current design choices can be improved upon. At the same time, more work is needed with different dictionary search and visualisation platforms, which can be catered also to the needs of specific user groups. One of the strengths of the current implementation is that we have a large amount of lexical data from different languages in the same infrastructure, and the new work is not disconnected from earlier efforts, but instead builds upon it. However, from the user perspective it is probably necessary to differentiate language and target group specific exports and views.

In building new systems, one has to always remember the importance of the longevity of the data. We recently got an important reminder of this as the servers of our service provider caught fire[8]. We take regular backups of the data of our system both as SQL dumps and as Giella XMLs. Backing the data up in the Giella XML format comes with the additional benefit of it being convertible into the ISO standardised TEI format (Rueter & Hämäläinen, 2019), which ensures that the lexicographic data remains readable even in the distant future.

We will also consider which is the best option for digital preservation of this work in some larger and more persistent infrastructure. Zenodo[9] is one obvious option to store versioned exports as well, and the exports that relate to individual published dictionaries should be stored also digitally with particular care, so that it is always possible to go back into individual versions. This is needed, for example, to quantify the changes between different dictionary editions. These questions are also strongly related to the dictionary editing workflows of the individual teams, although we believe that periodic publications and later improved editions is a model that remains relevant for many dictionary creators. With the online platforms, naturally, the question of release based updates and continuous updates also becomes important, and may vary from situation to situation.

The system has already been used by other researchers (Koponen & Kuokkala, 2021) to study Skolt Sami word derivation. This shows that the data stored in our dictionary

[8] https://web.archive.org/web/20210310232354/ https://www.ovh.com/world/news/press/cpl1787.fire-our-strasbourg-si
[9] https://zenodo.org/

system is also accessible for other researchers and it can be a useful resource in linguistic research. However, this has not been taken into account as a possible use case when developing the system. In the future, it would be important to conduct user studies in order to better understand the needs of linguistics researchers to better support their use cases.

The most important feature that needs to be further improved and adapted is the dictionary editing workflow that the user interacts with. It is especially important that this is done in close collaboration with the system users. This calls for identification of different usage patterns that the users have, including the documentation of various steps in the usage.

Adding a new lexical entry, editing relations, adding example sentences and searching for related entries are all tasks the dictionary editor will do continuously, and the interface should allow focused work where there are minimal interruptions and pauses caused by the underlying system. Ideally the information about usage bottlenecks would be collected by observing and tracking the real user actions in the interface, with their permission, and having continuous discussions about their experiences. However, it is particularly important to be able to distinguish the true obstacles in the editing platform, and issues that are related to insufficient training and documentation: a complex expert system will inevitably have some learning curve. From this point of view it is also important to distinguish the issues novice and expert users have, and to understand the process through which the novices become fully competent expert users.

Most of the dictionaries contain a large number of example sentences for each entry and meaning groups. Some of these are created by the dictionary editors, and some originate from various sources. The sources are in all cases important to indicate. When possible, the example sentences used in the dictionaries should be linked into different corpora and related datasets, both for accountability and the possibility to further provide access into them. This also makes it clear which materials, created by who, are actually used in the dictionaries, which makes citation of all sources used easier and benefits the visibility of previously done work in our scientific community. At the same time linked data also becomes more difficult to maintain when we cannot guarantee that all linked sources remain as accessible as our system.

Another area where similar connections could be created is multimedia. There are numerous spoken language corpora, some of which are openly licensed, and using their materials in connection with the dictionary resources would be an excellent addition, since our system doesn't currently have pronunciation information. It could be possible to add this information also in IPA or other transcription system, but in this day and age actual multimedia references seem very realistic and even expected.

5. Conclusions

In this paper, we have presented our open-source dictionary editing system that was developed for post-editing the new printed Finnish-Skolt Sami dictionary. We have described the system and how it interacts with the existing open-source language technology infrastructure called Giella. By releasing our source code openly on GitHub, we hope that other people can make use of our system to meet their dictionary editing needs.

We have developed the system taking into account the latest NLP tools available for Skolt Sami. This has made the dictionary editing process easier as automatically introduced information such as inflectional forms, derivations and compounds would have taken a great deal of time to annotate manually. The fact that our system makes it possible for the dictionary editors to fix errors in the automatically generated inflectional forms also benefits the development of the NLP tools used. Finally, we aimed at building a system that not only serves in producing a paper dictionary, but forces the editors to edit the lexicographic entries in such a way that they remain structured and parseable by computational means. This meant that the final dictionary was also easy to be made available online[10] in a searchable fashion.

6. References

Alnajjar, K., Hämäläinen, M., Rueter, J. & Partanen, N. (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*. Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), pp. 1–6. URL https://www.aclweb.org/anthology/2020.coling-demos.1.

Beesley, K.R. & Karttunen, L. (2003). *Finite-State Morphology*. Stanford, CA: CSLI Publications, pp. 451–454.

Bradley, J. & Skribnik, E. (2021). The many writing systems of Mansi: challenges in transcription and transliteration. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Gerstenberger, C., Partanen, N. & Rießler, M. (2017). Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 57–66.

Hämäläinen, M. (2021). Endangered Languages are not Low-Resourced! In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. RootRoo Ltd, pp. 1–11.

Hulden, M. & Silfverberg, M. (2021). The Principal Parts of Finnish Nominals. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Hunt, B., Chen, E., Schreiner, S.L. & Schwartz, L. (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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 122–126. URL https://www.aclweb.org/anthology/N19-4021.

Hämäläinen, M. (2019). UralicNLP: An NLP Library for Uralic Languages. *Journal of Open Source Software*, 4(37), p. 1345.

Hämäläinen, M., Partanen, N., Rueter, J. & Alnajjar, K. (2021). Neural Morphology Dataset and Models for Multiple Languages, from the Large to the Endangered. In *Proceedings of the the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa 2021)*.

[10] https://saan.oahpa.no/fin/sms/

Hämäläinen, M. & Rueter, J. (2018). Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In *Proceedings of the Eighteenth EURALEX International Congress*, pp. 967–978.

Koponen, E. & Kuokkala, J. (2021). Kantasaamen *-(ę)hčę-frekventatiivijohtimen edustuksesta nykyisissä saamelaiskielissä. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Lehtinen, M., Koponen, E., Fofonoff, M., Lehtola, R. & Rueter, J. (eds.) (2021). *Suomi–koltansaame-sanakirja Lää'dd-sää'm-sää'nnǩee'rjj*. Saamelaiskäräjät.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T.A. & Silfverberg, M. (2013). HFST a system for creating NLP tools. In *International Workshop on Systems and Frameworks for Computational Morphology*. Springer, pp. 53–71.

Littell, P., Pine, A. & Davis, H. (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*. Honolulu: Association for Computational Linguistics, pp. 141–150. URL https://www.aclweb.org/anthology/W17-0119.

Mechura, M. (2016). Data Structures in Lexicography: from Trees to Graphs. *RASLAN 2016 Recent Advances in Slavonic Natural Language Processing*, p. 97.

Měchura, M.B. et al. (2017). Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*. pp. 19–21.

Moseley, C. (ed.) (2010). *Atlas of the World's Languages in Danger*. UNESCO Publishing, 3rd edition. Online version: http://www.unesco.org/languages-atlas/.

Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T. & Tyers, F.M. (2014). Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. In *The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era"*. pp. 71–77.

Nasution, A.H., Murakami, Y. & Ishida, T. (2018). Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association. URL https://www.aclweb.org/anthology/L18-1536.

Pirinen, T. & Tyers, F. (2021). Building language technology infrastructures to support a collaborative approach to language resource building. In M. Hämäläinen, N. Partanen & K. Alnajjar (eds.) *Multilingual Facilitation*. Rootroo Ltd.

Pirinen, T.A. (2019). Building minority dependency treebanks, dictionaries and computational grammars at the same time—an experiment in Karelian treebanking. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*. pp. 132–136.

Rueter, J. & Hämäläinen, M. (2017). Synchronized Mediawiki Based Analyzer Dictionary Development. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*. pp. 1–7.

Rueter, J. & Hämäläinen, M. (2019). On XML-MediaWiki Resources, Endangered Languages and TEI Compatibility, Multilingual Dictionaries For Endangered Languages. In M. Gürlek, A. Çiçekler & Y. Taşdemir (eds.) *AsiaLex 2019*. Turkey: Asos Publisher.

Rueter, J. & Hämäläinen, M. (2020). FST Morphology for the Endangered Skolt Sami Language. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*. European Language Resources Association (ELRA).

Sammallahti, P. & Mosnikoff, J. (1991). *Suomi–koltansaame sanakirja*. GIRJEGIISA.
Wilbur, J. (2017). The Pite Saami lexicographic backbone From a FileMaker Pro database to published digital results. In Электронная письменность народов российской федерации: опыт, проблемы и перспективы, pp. 299–309.