On Editing Dictionaries for Uralic Languages in an Online Environment

Khalid Alnajjar
Department of
Computer Science
University of Helsinki
khalid.alnajjar
@helsinki.fi

Mika Hämäläinen
Department of
Digital Humanities
University of Helsinki
mika.hamalainen
@helsinki.fi

Jack Rueter
Department of
Digital Humanities
University of Helsinki
jack.rueter
@helsinki.fi

Abstract

We present an open online infrastructure for editing and visualization of dictionaries of different Uralic languages (e.g. Erzya, Moksha, Skolt Sami and Komi-Zyrian). Our infrastructure integrates fully into the existing Giellatekno one in terms of XML dictionaries and FST morphology. Our code is open source, and the system is being actively used in editing a Skolt Sami dictionary set to be published in 2020.

Abstract

Tämä artikkeli esittelee Uralilaisten kielten (kuten ersän, mokshan, koltansaamen ja komi-syrjäänin) sanakirjojen toimittamiseen ja visualisointiin tarkoitetun avoimen verkkoinfrastruktuurin. Meidän infrastruktuurimme integroituu Giellateknoon XML-sanakirjojen ja FST-morfologian osalta. Lähdekoodimme on avointa, ja järjestelmäämme käytetään tällä hetkellä aktiivisesti koltansaamen sanakirjan toimitustyössä. Koltan sanakirja julkaistaan vuonna 2020.

1 Introduction

In order to revitalize severely endangered languages, such as many of the Uralic languages, enormous work is required to collect as many resources and knowledge about them as possible, while also involving their native communities. Digitizing the resources of endangered languages is crucial as it boosts the language resources in various ways, such as preserving them in a versioned manner and facilitating access to them globally. Scholars have produced valuable lexicographic resources (such as dictionaries and finite-state transducers) for endangered Uralic languages (e.g. Komi-Zyrian, Ingrian, Erzya, Moksha and Skolt Sami) in order to revitalize them.

We present a large-scale open-source MediaWiki-based dictionary for such languages, (named Akusanat) (c.f. Hämäläinen and Rueter 2018) and a customly-built and user-friendly web system (named Ve′rdd[1]) that improves and amending the knowledge presented in such dictionaries. As MediaWiki sets some limitations to the structure of the system both on the back-end in terms of the database and on the front-end in terms of usability, the external, yet integrated system, Ver′dd is set to tackle these limitations.

It is also worth noting that one use case where such dictionary interfaces could be very useful is language documentation. Although the field has been slow to adopt language technology, there have been significant recent advancements in integrating it into projects workflows (Blokland et al., 2015; Блокланд et al., 2014). One aspect where this has not yet been done is lexicography, which, however, is usually considered a central part of language documentation efforts. The field is still largely dominated by aging and poor software which clearly is not easily compatible with modern needs. Our system cannot be used offline, which is a challenge for language documentation use, but it could very well finds its place as an easy shared interface between researchers and community members.

2 Related Work

There is a myriad of active online dictionary projects targeting only one language that are under development by different people, who oftentimes are unaware of each other's contributions. In this section, we present some of the recent work on online dictionaries, which is heavily guided by the needs of one individual language. Our infrastruc-

[1]Skolt word for stream

Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages, pages 26–30
Wien, Austria, January 10 — 11, 2020.
https://doi.org/10.18653/v1/FIXME

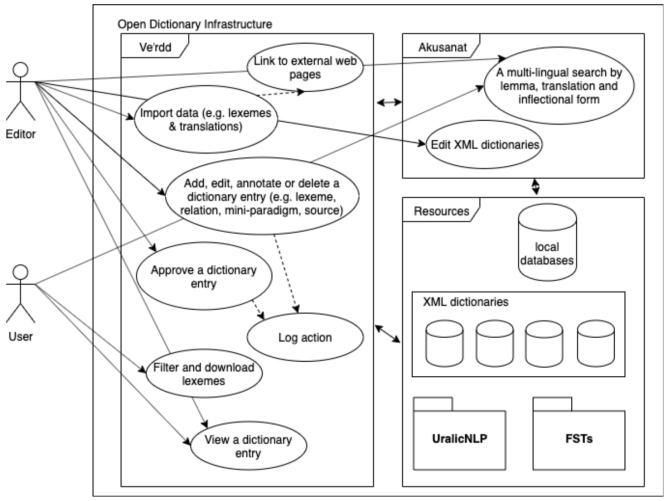

Figure 1: A UML diagram illustrating use-cases of the infrastructure

ture differs from these projects in that its driving design principle is multilinguality and support for a multitude of different Uralic languages.

A recent dictionary for St. Lawrence Island Yupik (Hunt et al., 2019) combines Foma-based morphological analyzers with an HTML based search interface. Unlike Akusanat, which does the morphological analysis and generation in the cloud, their solution runs the transducers on the client side with Foma's Javascript integration.

The Livonian dictionary consists of three databases, one – lexical, the second – morphological, and the third – a text corpus. While lemmas and their data are stored in the lexical database, and morphological forms are documented in the morphological database, all words indexed in the corpus refer to lemmas in the lexical database. Thus, all materials in the cluster can be accessed directly from the three databases (c.f. Ernštreits 2019).

There are also various attempts to build infrastructure for national majority languages. These projects also seem to be characterized by simultaneous use of different tools, with various connections to commercial software providers (see Tavast et al. 2018). Also from this point of view there is clear demand for open and easily customizable dictionary editing and data retrieval platforms, such as the infrastructure presented here.

3 The Open Dictionary Infrastructure

Akusanat is built using MediaWiki. MediaWiki is a well documented and open-source framework that comes with a set of fulfilled quality attributes such as support for multiple simultaneous users, user account management and a documented API. In addition, MediaWiki has been perceived as a useful framework for dictionaries in the past (Laxström and Kanner, 2015).

Despite the features that MediaWiki has, it does not provide an intuitive editing interface. This hinders the involvement of users of non-technical backgrounds, which is often the case for many native speakers of endangered languages. As a result, involving the native community in improving and approving the recorded information in the dictionaries is not possible. Ve′rdd is built to tackle this issue while granting users and language experts the ability to contribute to different aspects of the knowledge of such endangered languages. Additionally, Ve′rdd makes different and scattered lexicographic resources in the system available for researchers and non-academic dictionary users alike. Figure 1 shows the infrastructure of our open dictionary on a high-level of abstraction showing how different users can interact with it, revealing the interplay of the two systems: Ve′rdd and Akusanat.

3.1 Akusanat

The Akusanat dictionaries offer a distinct presentation of synchronized data shared with the Giella (Giellatekno, Divvun) infrastructure. Like the Giella dictionaries (Moshagen et al., 2014), Akusanat utilizes HFST-based (Lindén et al., 2013) finite-state transducers but with an open-source python library (UralicNLP (Hämäläinen, 2019)) in the search field, which allows users the option of entering virtually any word form to locate a pos-

sible lemma. Unlike the Giella dictionaries, however, Akusanat provides language internal links to associate words with derivational stems as well as external links to translations and cognates in other language dictionaries within Akusanat and entirely independent databases outside the domain.

The lexicographic data of Akusanat originates from the XML-based dictionaries in the Giellatekno infrastructure. Akusanat provides a user-friendly way of accessing the lexicographic data both as a regular dictionary user and as a dictionary editor solving the XML bottleneck. This means that, unlike XML, the lexicographic data can be edited simultaneously by multiple users. All the edits done in the Mediawiki-based Akusanat environment are synchronized with the XMLs residing in the Giellatekno infrastructure (c.f. Hämäläinen and Rueter 2019). However, at the same time, also editing of original XML files is possible, as the synchronization works to both directions.

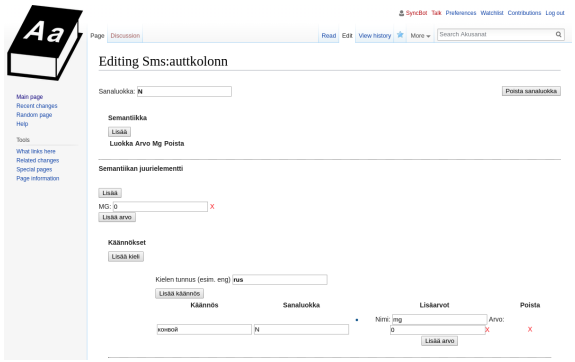

Figure 2: Edit form on Akusanat

Akusanat benefits greatly from the inbuilt quality attributes[2] of MediaWiki, such as user management, admin view, a documented Wiki-syntax and an open MediaWiki API. However, MediaWiki comes with a multitude of limitations; first and foremost, Akusanat requires a web application separate from MediaWiki to handle the synchronization of the dictionary data between the XMLs and the MediaWiki database. In addition, usability is limited by what MediaWiki has been developed for. By default, MediaWiki exposes the full Wiki syntax of each page for editing. In a dictionary setting, where the integrity of the data structure needs to be ensured, such free editing functionality has to be limited. This has been solved by introducing an edit form as seen in Figure 2. However, the more complex the demands for the system become in terms of editing, search etc., the more challenging it becomes to integrate the desired features to MediaWiki as opposed to developing a new system from scratch.

[2]For more discussion on quality attributes on web applications, see Offutt 2002

3.2 Ve′rdd

Ve′rdd is a Django-based custom developed system. The use of Django as a framework can be motivated by the fact that it scores high when compared to other popular web frameworks (Plekhanova, 2009). The goal of Ver′rdd is to correct the shortcomings of Akusanat on the intuitiveness of editing, since Akusanat users must be familiar with the structure of the XML dictionaries while editing the lexicographic entries. Ve′rdd stores information in an SQL database isolated from Akusanat which gives trusted editors the ability to perform amendments to information present in it without interfering with online dictionaries in Akusanat. Whereas Akusanat is meant to present an openly available bleeding edge version of the dictionaries, Ver′rdd, on the other hand, is tailored towards a more curated dictionary editing without immediately exposing all the edits to online users.

User experiences based on interactions with the system are continuously taken into account to facilitate the usability of the system and provide non-technical and technical users robust means for accessing and improving knowledge present in the database. Currently, the system is in use by dictionary editors authoring a Finnish-Skolt Sami dictionary and verifying the entries in it with the aim of publishing an online and a printed dictionary in early 2020. The needs of these non-technical users have been and are continuously being taken into account in the development of Ve′rdd.

Figure 1 lists the core interactions of common users (speakers or learners of the endangered language) and editors with the system. The system supports import from XML dictionaries and CSV files. Whenever data is imported, Ve′rdd consults multiple resources (e.g. Akusanat, UralicNLP and FSTs) to retrieve missing information such as part-of-speech, continuation lexica and mini-paradigms which ensures that imported information contains all the details present in other systems. Users and editors can then filter and order lexemes using multiple criteria (such as language, consonance, etc.) as seen in Figure 3.

By using Ve′rdd, editors have the ability to modify and comment on any present information in the

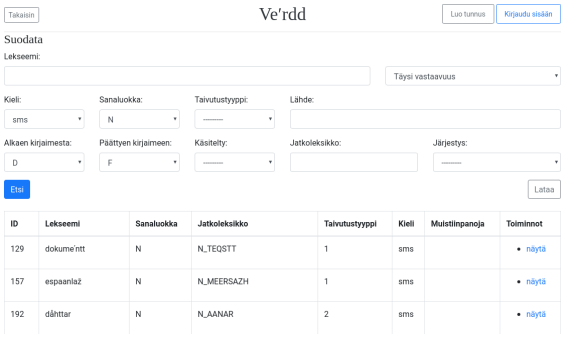

Figure 3: Search interface on Ve′rdd

database. To encourage the involvement of native speakers of endangered languages, especially speakers of another non-endangered language such as Russian or Finnish, the system allows approved editors with such criteria to add, edit, comment on and confirm the knowledge presented in the database. This guarantees that the information present in the system is validated and accurate as opposed to Akusanat, in which anyone can create an account and make edits. Whenever an editor performs any action (e.g. adding a lexeme or a translation), the system keeps a log which allows discovering cases of conflict and reverting back in the case of incorrect or non-verified actions are applied.

In Ve′rdd, all lexemes are stored as independent entities in the database. These independent entities are linked to each other with an abstract notion of relation. A relation between two lexemes has a direction and it can contain additional information. The system currently supports a multitude of relations such as translation, cognate or derivation. Derivational information is automatically gathered from the FSTs when data is imported to the system. An example of the relations view is show in Figure 4, the relations can be modified in their respective edit interface.

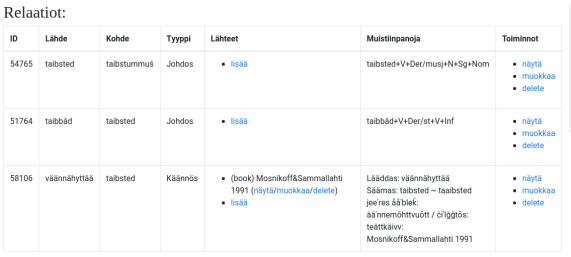

Figure 4: Relations for the word *taibsted* in Ve′rdd

For general editing, Ve′rdd exposes only the relevant information to the editor in an interface that is more narrowed down than the full-blown complexity of Akusanat edit form. Ve′rdd highlights only the essential for the dictionary editor for the particular task. This is why relations, lexemes and morphologies are edited in different interfaces; all of them accessible from the general view on the lexeme. The lexeme level editing interface is seen in Figure 5.

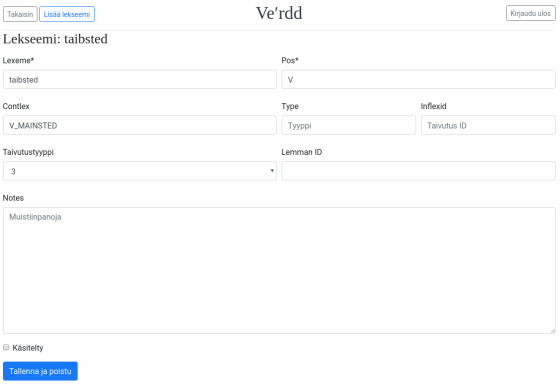

Figure 5: General edit interface for lexemes in Ve′rdd

In a timely manner, Ve′rdd can then send the approved information (by authorized experts and native speakers) to Akusanat and other resources (e.g. UralicNLP and FSTs), which would then make retaining up-to-date information across multiple resources possible; hence, reducing the risk of providing inaccurate and misleading information.

4 Discussion and Conclusions

Ve′rdd is already being used by the Skolt Sami dictionary editors as this is being written. Our development strategy involves direct interaction between the actual end users and designers, which has helped to address issues and features foreseen at the onset. A later goal would be to integrate Ve′rdd and Akusanat more completely into the infrastructure where morphological analysers and other tools are being used, so that the end-user would have a natural and intuitive environment to work with the lexicon, but so that these changes would be automatically included into the newest compiler analyzer.

More work should also be done in connecting the lexicographic resources into various corpora that are openly available. There are various ways to proceed with this: the examples could be extracted automatically, the examples could be selected with references to the corpora, or the corpora could be tagged for representative examples that would be picked into dictionary.

The most important goal in the further development of Ve′rdd must, however, be further collaboration with the users. The system will be continuously improved with the received feedback, and the user base has to be widened to encompass a larger number of users in different languages included in the project.

The source code of the systems has been made available on Bitbucket[3].

References

Rogier Blokland, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. Language documentation meets language technology. In Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, editors, *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway*, number 2015:2 in Septentrio Conference Series, pages 8–18. The University Library of Tromsø.

Valts Ernštreits. 2019. Lexical tools for low-resource languages: A livonian case-study. *Electronic lexicography in the 21st century (eLex 2019): Smart lexicography*, page 103.

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. 2019. UralicNLP: An NLP library for Uralic languages. *Journal of Open Source Software*, 4(37):1345.

Mika Hämäläinen and Jack Rueter. 2018. Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In *Proceedings of the Eighteenth EURALEX International Congress*, pages 967–978.

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

Niklas Laxström and Antti Kanner. 2015. Multilingual semantic mediawiki for finno-ugric dictionaries. In *Septentrio Conference Series*, 2, pages 75–86.

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 *International workshop on systems and frameworks for computational morphology*, pages 53–71. Springer.

[3]https://bitbucket.org/mokha/verdd/src/master/ and https://bitbucket.org/mikahama/saame/src/master/

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 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"*, pages 71–77.

Jeff Offutt. 2002. Quality attributes of web software applications. *IEEE software*, 19(2):25–32.

Julia Plekhanova. 2009. Evaluating web development frameworks: Django, ruby on rails and cakephp. *Institute for Business and Information Technology*.

Arvi Tavast, Margit Langemets, Jelena Kallas, and Kristina Koppel. 2018. Unified data modelling for presenting lexical data: The case of ekilex. In *Ed. J. Čibej, V. Gorjanc, I. Kosem & Simon Krek, Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts, Ljubljana*, pages 749–761.

Рохир Блокланд, Михаэль Рийсслер [Рисслер], Нико Партанен, Марина Федина, and Андрей Чемышев. 2014. Использование цифровых корпусов и компьютерных програм в диалектологических исследованиях. pages 252–255. ИИЯЛ УНЦ РАН.