HELDA
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

The Open Dictionary Infrastructure for Uralic Languages

Alnajjar, Khalid

þÿ 0 H : 8 @ A : 0 O M = F 8 : ; > ? 5 4 8 O
2019

Alnajjar , K , Hämäläinen , M , Partanen , N & Rueter , J 2019 , The Open Dictionary
þÿ I n f r a s t r u c t u r e f o r U r a l i c L a n g u a g e s . i n - ; 5 : B @ > = = 0 O 8 A L < 5 = = > A B L
þÿ $ 5 4 5 @ 0 F 8 8 : ? K B , @ > 1 ; 5 < K 5 @ A ? 5 : B 8 2 K . 0 H : 8 @ A : 0 O M = F 8 : ; > ? 5 4
þÿ 4 9 - 5 1 , - ; 5 : B @ > = = 0 O ? 8 A L < 5 = = > A B L = 0 @ > 4 > 2 > A A 8 9 A : > 9 $ 5 4 5 @ 0 F 8 8
þÿ ? 5 @ A ? 5 : B 8 2 K , U f a , R u s s i a n F e d e r a t i o n , 2 7 / 1 1 / 2 0 1 9 .

http://hdl.handle.net/10138/308789

unspecified
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

ЛИНГВИСТИЧЕСКИЕ БАЗЫ ДАННЫХ И ПРИЛОЖЕНИЯ. ФУНКЦИОНИРОВАНИЕ ЯЗЫКА В СЕТИ ИНТЕРНЕТ

Х. Альнайар, М. Хаммалайнен,
Н. Партанен, Дж. Рюйтер
г. Хельсинки, Финляндия

ИНФРАСТРУКТУРА ОТКРЫТЫХ СЛОВАРЕЙ УРАЛЬСКИХ ЯЗЫКОВ

Мы представляем открытую онлайн-инфраструктуру для редактирования и визуализации словарей разных уральских языков (например, эрзя, мокша, скольт-саамский и коми-зырянский). Наша инфраструктура полностью интегрируется в существующую Giellatekno с точки зрения словарей XML и морфологии FST. Наш код в открытом источнике.

Ключевые слова: уральские языки, онлайн словари, словари в формате XML.

K. Alnajjar, M. Hämäläinen,
N. Partanen, J. Rueter
Helsinki, Finland

THE OPEN DICTIONARY INFRASTRUCTURE FOR URALIC LANGUAGES

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.

Keywords: Uralic languages, online dictionary, XML dictionaries

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) [Hämäläinen, Rueter 2018] and a customly-built and user-friendly web system (named Ve′rdd) that improves and amending the knowledge presented in such dictionaries.

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 infrastructure 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 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 [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 [Tavast 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.

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, 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.

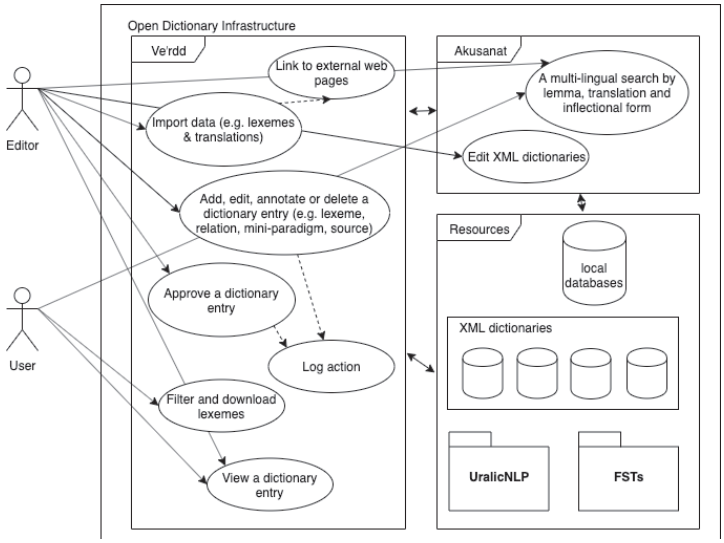

Figure 1. A UML diagram illustrating use-cases of the infrastructure

The Akusanat dictionaries offer a distinct presentation of synchronized data shared with the Giella (Giellatekno, Divvun) infrastructure. Like the Giella dictionaries [Moshagen 2013; Rueter 2013], 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 possible 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 [Hämäläinen, Rueter 2019]. However, at the same time, also editing of original XML files is possible, as the synchronization works to both directions.

Ve′rdd is a customly developed system that fixes the shortcoming of Akusanat on the intuity of editing, since Akusanat users must be familiar with the editing structure of the XML dictionaries. Ve′rdd stores information in an isolated database from Akusanat, which gives trusted editors the ability to perform amendments to information present in it without interfering with online dictionaries in Akusanat. Also, 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 of accessing and improving knowledge present in the database.

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 by multiple criteria (such as their language, consonance, etc).

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

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 analyzers 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.

References

Ernštreits V. (2019). Lexical tools for low-resource languages: A Livonian case study. In Proceedings of the eLex 2019 conference. P. 161–176.

Hämäläinen M., & Rueter, J. M. (2019). An Open Online Dictionary for Endangered Uralic Languages. In Proceedings of the eLex 2019 conference. P. 819-830.

Hämäläinen M., & Rueter J. M. (2018). Advances in synchronized XML-MediaWiki dictionary development in the context of endangered Uralic languages. In Proceedings of the XVIII EURALEX. P. 967–978.

Hämäläinen M. (2019). UralicNLP: An NLP Library for Uralic Languages. Journal of open source software, 4(37), [1345].

Hunt B., Chen E., Schreiner, S. L. R., & Schwartz L. (2019). Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In Proceedings of NAACL 2019.

Laxström N., & Kanner A. (2015) Multilingual Semantic MediaWiki for Finno-Ugric dictionaries. In Septentrio Conference Series P. 75-86.

Lindén K., Axelson E., Drobac S., Hardwick S., Kuokkala J., Niemi J., Pirinen T., et al. (2013). HFST a system for creating NLP tools. In International workshop on systems and frameworks for computational morphology P. 53–71.

Moshagen S. N., Pirinen T., & Trosterud T. (2013). Building an open-source development infrastructure for language technology projects. In Proceedings of the 19th Nordic Conference of Computational Linguistics/ P. 343–352.

Tavast A., Langemets M., Kallas J., and Koppel K. (2018) Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In Proceedings of the XVIII EURALEX. P. 749-761.