Advances in synchronized XML-media wiki dictionary development in the context of endangered uralic languages

Hämäläinen, M.

Ljubljana University Press
2018

Hämäläinen , M & Rueter , J 2018 , Advances in synchronized XML-media wiki dictionary þÿdevelopment in the context of endangered uralic languages . in J ibej & S Krek (eds) , Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts : 17-21 July 2018, Ljubljana . EURALEX Proceedings , Ljubljana University Press , Ljubljana , pp. 967-978 , EURALEX International Congress , Ljubljana , Slovenia , 17/07/2018 . < https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1 >

http://hdl.handle.net/10138/313234

Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages

Mika Hämäläinen, Jack Rueter
Department of Digital Humanities, University of Helsinki
E-mail: mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi

Abstract

We present our ongoing development of a synchronized XML-MediaWiki dictionary to solve the problem of XML dictionaries in the context of small Uralic languages. XML is good at representing structured data, but it does not fare well in a situation where multiple users are editing the dictionary simultaneously. Furthermore, XML is overly complicated for non-technical users due to its strict syntax that has to be maintained valid at all times. Our system solves these problems by making a synchronized editing of the same dictionary data possible both in a MediaWiki environment and XML files in an easy fashion. In addition, we describe how the dictionary knowledge in the MediaWiki-based dictionary can be enhanced by an additional Semantic MediaWiki layer for more effective searches in the data. In addition, an API access to the lexical information in the dictionary and morphological tools in the form of an open source Python library is presented.

Keywords: online dictionary, collaborative editing of XML, Semantic MediaWiki dictionary

1 Introduction

In this paper, we present advances in the development of our open-source synchronized XML-MediaWiki dictionary environment[1] (Rueter & Hämäläinen 2017). The dictionary data consists of multiple XML dictionaries for small Uralic languages[2] following the same XML structure. XML dictionaries are used on the Giellatekno infrastructure (Trosterud, Moshagen & Pirinen 2013) for many distinct facets of linguistic research such as Intelligent Computer-Assisted Language Learning (ICALL) (Antonsen et al. 2014), FST generation for morphological analyzers and spellcheckers.

XML is a great format for storing structural data, such as information usually stored in a dictionary. It does, however, have some drawbacks, such as editing XML data in a collaborative fashion is a challenging task. This is even more so in the case of non-technical people native in endangered Uralic languages. In order to enable them to produce and correct dictionary resources, a simplified way to edit XML data is needed.

We have thus developed a MediaWiki-based online dictionary system, the purpose of which is to make it possible to edit structural dictionary data collaboratively with a simplified interface. The dictionary works in such a way that we can get the edits instantly in our XML formalism, and edits made directly in the XMLs are also updated to the MediaWiki.

1 Available on https://sanat.csc.fi/

2 The languages currently supported in the dictionary are Skolt Sami, Ingrian, Meadow Mari, Votic, Olonets-Karelian, Erzya, Moksha, Hill Mari, Udmurt, Tundra Nenets and Komi-Permyak

2 Related Work

This section presents some of the previous research done in the context of online dictionaries. The previous work ranges from theoretical takes on online dictionaries to actual online systems implemented for the task. In a meta-analysis of studies on the usage of electronical dictionaries (Töpel 2014), several advantages in electronical dictionaries were identified. A positive impact on speed, performance, ease of use, vocabulary retention and satisfaction were reported in a dictionary use situation.

The XML structures of this project are compatible with and, where possible, identical to those used by the dictionaries in the Giellatekno infrastructure, where local enhancement provides the availability of special glyphs for assistance in individual language input, links to corpora search in Giellatekno-hosted Korp, as well as grammatical links for the enlightenment of the lay user[3]. Whereas the Giellatekno dictionaries provide for dictionary users without specific keyboards for the individual languages, we require our users to have keyboards of their own[4]. Pointer data in our XML and MediaWiki interfaces allow us to open individual page links to the etymological database for Sami languages (Álgu-tietokanta 2002).

Uralic language databases are the target of continuous development in Estonia. This can be observed in the outline of Estonian and Uralic language archive materials in Tallinn and Tartu (Viikberg 2008), and subsequent mention of work on Estonian-Mari and Estonian-Erzya dictionaries at EKI (Eesti Keele Instituut [Estonian Language Institute]) in EELEX (Tender et al. 2017). Similar bilingual dictionary development with audio resources are described for Võro-Estonian (Männamaa & Iva 2015).

The Dictionary of Old Norse Prose (ONP) (Johannsson & Battista 2017) has implemented multiple search and presentation features. It strives towards an online tool with enhanced corpus search and allows for presentation of manuscript and archive materials, as well as individuated download possibilities. In our project, however, we retain a light structure with synchronic editing for XML and MediaWiki. The XMLs contain a set of hand-selected example sentences from corpora to be displayed to the user in the online dictionary. However, our system has not been linked to full corpora for example sentence extraction.

The role of e-lexicography is growing. Not only is the detail required for the conversion from printed dictionaries to digital format being examined, but investigations are also being made of the feasible saturation of data presentation. E-lexicography allows for the introduction of new tools, and is seen as an opportunity to provide direct data extraction from various data sources (Bothma, Gouws & Prinsloo 2017). Our MediaWiki presentation involves three dimensions of linking. It includes links to external datasets (etymology and audio), other languages in the internal dataset (definitions, etymology), and dictionary internal links between articles (compound word constituents and derivation stems). We also generate regular paradigmatic tables for viewing while retaining a view of lemma, native definition, translation and morphologically important category information on the screen.

3 Giellatekno online morphologically savvy dictionaries with click-in-text readers and possible Korp links are available at: http://sanit.oahpa.no/ (North Sami), http://baakoeh.oahpa.no/ (South Sami), http://saanih.oahpa.no/ (Inari Sami), http://saan.oahpa.no/ (Skolt Sami), http://sanat.oahpa.no/ (Northern Balto-Finnic languages), http://sonad.oahpa.no/ (Southern Balto-Finnic languages), http://valks.oahpa.no/ (Mordvin languages), http://muter.oahpa.no/ (Mari languages), http://kyv.oahpa.no/ (Permic languages), and http://vada.oahpa.no/ (Nenets).

4 The necessary keyboards for most Uralic languages are produced for Windows, Mac and Android and available at http://divvun.no/ for Saamic languages, and analogical keyboards for other languages can be generated directly in the Giellatekno infrastructure.

3 The XML Dictionaries

The XML dictionaries draw upon the goal of minimizing data redundancy in different branches of an extended infrastructure at Giellatekno (Trosterud, Moshagen & Pirinen 2013). Original parallel sources existing for online morphologically savvy translation dictionaries, on the one hand, and minimal sized ICALL dictionaries, on the other, have been integrated with lemma:stem pair data utilized in transducer production. Subsequently, other research data has been incorporated into the XML structure as well, such as audio pointers, and etymological as well as derivational information partially inherited from previous language projects. Thus, while the dictionaries can be used through XSL transformation to provide code output (lexc) for the construction of transducers used in finite-state morphological analyzers and spell checkers, they also serve as extensive databases for other research projects. The distinction between source and target languages is maintained utilizing ISO 639-3 three-letter codes, which can be attested in the XML root element as well as the translation group <tg/>, example group <xg/>, etymon and cognate elements.

The translation dictionaries were originally set up as source-to-target, bi-lingual dictionaries. In word entries with broader semantic coverage, granularity has been introduced. This allows for multiple translations in the instance of semantically close definitions <t/>, and separate meaning groups <mg/> for distinct senses of a word. Contextual usage is demonstrated in Figure 1, which shows translation groups within the semantically appropriate meaning group. The Giellatekno dictionaries based on this XML structure are available and undergoing continuous development within the Giellatekno infrastructure.

```
<e>
  <rev-sort_key>issaa</rev-sort_key>
  <lg>
    <l pos="N">aassi</l>
    <etymology>
      <etymon algu_lekseemi_id="82224" id="246450" xml:lang="sms">aassi</etymon>
    </etymology>
    <stg>
      <st Contlex="N_PRSPRC-VVKK-I">aassi</st>
    </stg>
    <inc-audio>
      <c name="ID_Audio">3341</c>
    </inc-audio>
    <compg drv="V»N" type="Der">
      <comp drv="" ord="E2" pos="Suf">Der/NomAg</comp>
    </compg>
  </lg>
  <mg relId="0">
    <tg xml:lang="eng">
      <t pos="N">resident</t>
    </tg>
    <tg xml:lang="rus">
      <t pos="N">житель</t>
    </tg>
    <xg>
      <x src="JS2 06749_2az 0:02:36">To´b lij poostai päi´ǩǩ, jeä´la ni ǩeäk jeänaš aazzi.</x>
      <xt xml:lang="fin">Se on syrjäseutua, ei ole juuri ketään asukkaita.</xt>
    </xg>
  </mg>
</e>
```

Figure 1: XML entry

Optional enhancement of the underlying lemma (e/lg/l), stem (e/lg/stg/st) and inflection (e/lg/stg/st@Contlex) dictionaries can be observed in the etymon and audio pointers, as well as the derivation (e/lg/compg), translation (e/mg/tg) and example (e/mg/xg) groups. While the lemma, stem and continuation lexica data serve as vital information in tranducer development, etymology, audio and compounding pointers provide for navigation between and within dictionaries. The etymon pointer

is used to access an online Sami language etymology dictionary, whereas an optional cognate sibling allows for pointing between languages in the MediaWiki infrastructure. Likewise, the inc-audio/audio pointers allow for accessing recordings in the Max Planck Institute archives at Nijmegen, and compounding group pointers offer access for navigation within the source language at the lemma and suffix levels.

Homonymy is addressed on a part-of-speech basis, with words bearing mutual etymological and inflectional data subordinated to single entries but feasibly different senses.

4 The Synchronized Dictionary System

The synchronized dictionary system we are proposing is meant to solve the problem of collaborative editing of XML dictionaries. Having multiple editors modifying the contents of pure XML dictionaries simultaneously is not an easy task to accomplish. It gets even more difficult if the editors have only a very limited technical background and from little to no understanding of the XML syntax. Large-scale tasks such as crowdsourcing of dictionary editing become next to impossible with plain XML files.

Another XML specific limitation our system is made to solve is breaking the tree structure of XML. Our dictionary system can build links in between different lexical entries even across multiple dictionaries to provide a more graph like structure of the dictionary data. This also makes it possible to conduct more complex queries to search for information stored in the dictionary system.

What makes our system synchronized is that we do not want to move entirely away from the XML standard, but rather build a system in which the same dictionary information can be edited in an easier crowdsourced fashion in a MediaWiki environment and also directly in the XMLs, so that edits at either end of the system will be made instantly available to all viewers of the dictionary system.

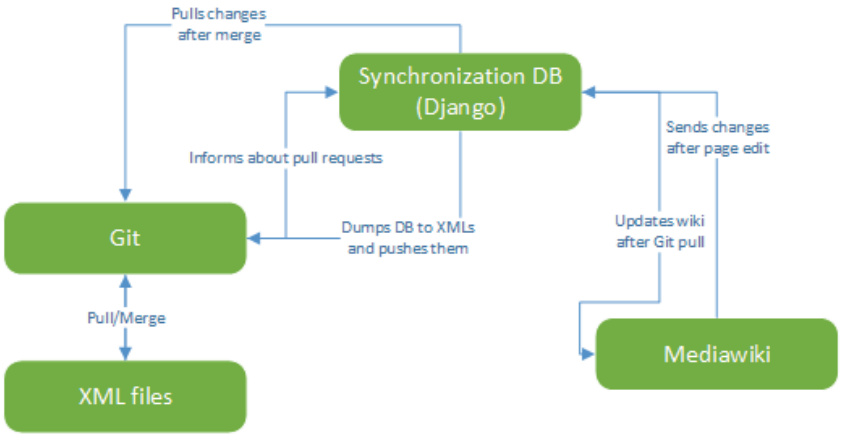

Figure 2: System architecture

The core of the dictionary system is the Django-based Synchronization DB seen in the middle of Figure 2. This database application provides APIs to the Git integration for making changes to the XML files and also APIs to the MediaWiki integration to communicate with the Wiki environment. The different parts of the system are discussed in more detail in the following subsections.

The non-functional requirements for the system are reliability and scalability. The dictionary system should not only serve researchers but also any language user outside of academia, which is a reason

why reliability of the platform is needed to guarantee a decent uptime. One of the design principles is that we should be able to include new dictionaries in the system, which means that it should scale well. The system should also be built fully on open source technologies in order to ensure its compatibility and maintainability in the future. Because the underlying MediaWiki platform and the XMLs will be used for other purposes than those required by our system, we also need to follow the idea of separation of concerns in order to fill a criterion of integratability.

4.1 Synchronization Database

The role of the synchronization database is to keep the most up-to-date version of the data in all situations. This makes it possible to isolate the synchronization feature from the XMLs and the MediaWiki, making it possible to introduce new sources and views to the data in the future. These might be XMLs following a different structure or an entirely new system for collaborative editing. By embracing the notion of separation of concerns, we do not want to build the synchronization database to follow the structure of MediaWiki syntax or XML syntax, but rather we want it to have its own scalable structure.

When constructing the system, we want to keep the option open for introducing new data sources, in XML or in another format. This means that the contents of the data are not predictable, and thus defining an SQL database would make incorporating new kinds of data difficult. Storing data in plain XML format is not a viable option either, as using an XML database has a huge negative impact on the performance of the system (Nicola & John 2003).

We thus propose using MongoDB as a solution for storing the data in an effective fashion. MongoDB is a so-called NO-SQL database which does not require a predefined structure for the database. In performance terms, it can run faster than a traditional SQL database (Boicea, Radulescu & Agapin 2012), making it a good option for our purposes.

The database application is a Django-based web application. Communication from the Git side and MediaWiki side with the database is thus done by using HTTP requests to the web application API. The process of bringing XML files to the system is done over Git.

When XMLs are edited or brought for the first time to the system, the changes made in them are committed to a Git repository. If the XML dictionaries already exist in the system, the editor has to run a special command line script that will create a new branch and dump the data from the synchronization database into XML format in that branch. This leaves the conflict resolution to the editor of the XML files. He can compare his current working branch with the latest data in the synchronization database, resolve the possible conflicts and merge the branches to the master branch. When the repository is pushed, the synchronization database pulls the changes and updates its internal database after which it starts updating the MediaWiki side.

The XMLs are read into the internal JSON format of the system by language and/or XML structure specific modules. When the XMLs are requested from the system, a format and language specific Django template is used to produce the XML structure. This conversion process of the data is explained in more detail in the next section.

4.2 Support for Multiple Languages

The system is built in a modular way to facilitate the inclusion of new languages or data sources. At the moment, all of the languages in the system follow the Giellatekno XML syntax, which means that the same modules are reused just with a different language flag. The system needs two language modules, one to handle the XML to JSON conversion for MongoDB and another to handle the JSON

to MediaWiki syntax conversion. We are dealing with languages whose orthographies contain special characters. This means, for multiple language support, that we have made sure the data is handled in UTF-8 format in all parts of the process.

Since the synchronization database itself is unaware of the contents of the data, how the XML gets transformed into JSON can be decided for each language module separately to better suit the needs of each dictionary type. The module currently developed for the Giellatekno XML does quite a direct transform of the XML data into JSON format. We do, however, handle homonyms differently in the JSON. In the Giellatekno format, homonyms are completely separate entries in the XML with a *hid* attribute to indicate the ID of the homonym. In the JSON format, it will be noted, we include all homonyms in a list under the main entry, which is identified by the lemma. The reason for this is simple: on the MediaWiki side all the homonyms are listed inside the same article which is identified by the lemma. Having all the different homonyms in the same entry in the synchronization database makes producing a MediaWiki page much simpler.

The other part of the language module is a script that can be run both in the synchronization system side and in the MediaWiki side to do a conversion between JSON and MediaWiki syntax. The important part is that the MediaWiki syntax is only used for the visualization of the dictionary data. For editing the dictionary entries in the MediaWiki side, a dump of the JSON data is included in the article in a hidden div element.

4.3 MediaWiki Integration

The MediaWiki integration is an extension which is isolated to work with a predefined set of namespaces. Our system creates a new MediaWiki namespace for each language. In practice, this means that each entry is prefixed by a three letter ISO language code, for example the Skolt Sami word *sokk* is stored inside of the MediaWiki article named *Sms:sokk*. The reason why it is important to limit the functionality with namespaces is not only that the namespace tells which language module should be used, but also that our dictionary system is a part of a shared MediaWiki dictionary of the Language Bank of Finland with multiple different data providers. This additional namespace restriction makes sure that our solution does not interfere with the MediaWiki entries other projects are building.

The MediaWiki extension of our system, in addition to communicating the changes to the synchronization database, provides the functionality for two MediaWiki article views: visualization and editing. A language specific module is used to construct a viewable version of a dictionary entry, or an article in the MediaWiki terminology. As described before, this viewable version stores the JSON structure as a hidden element for editing purposes.

The editing part of the MediaWiki extension solves the problem of XMLs requiring additional technical knowledge to be edited. The edit view of a MediaWiki article hides the MediaWiki syntax editor that would be shown by a MediaWiki based system by default. Instead, the editor constructs a form based on the language module and the hidden JSON element as seen in Figure 3. When we force users to edit the data through a form, we can make sure that the data is in a valid, parseable format. There is thus no possibility for the user to accidentally break the syntax of the data structure by, for example, forgetting a closing tag. Additionally, using a form for editing makes it possible for us to do form validation before saving the data in the system. At the moment, the validation means removing empty entries, such as a language entry without any translations.

Saving the edit form makes the system update the hidden JSON element and reconstruct the edit view based on the new JSON data using the exact same functionality as when a synchronization database pushes a JSON entry to the MediaWiki side. New changes are then immediately communicated to the synchronization database through the MediaWiki extension.

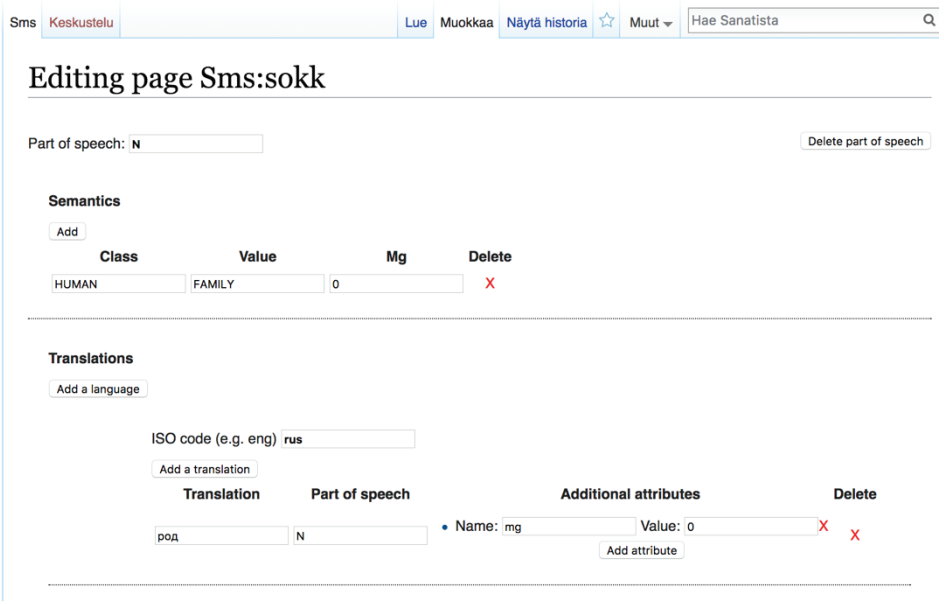

Figure 3: Form in MediaWiki

Since the MediaWiki stores each dictionary entry as a separate article, and the synchronization database does a similar separation, collaborative editing is made possible. Changes can be communicated between the two systems per entry basis without the need to parse an entire collection of lemmas, as in the case of XML. This structural separation of entries means that if different dictionary entries are edited simultaneously, there will not be any conflicts, but multiple edits can be synchronized in real time. The only case of simultaneous editing that is not supported is when the same MediaWiki article is edited at the same time by multiple users.

In addition to editing and visualizing the data, the MediaWiki integration has a search functionality for accessing the dictionaries. This is needed because the MediaWiki environment contains so many different dictionaries and word lists that using the default search box provided by MediaWiki makes it next to impossible to find the words in the system for an average user who is not familiar with the namespacing used in the system.

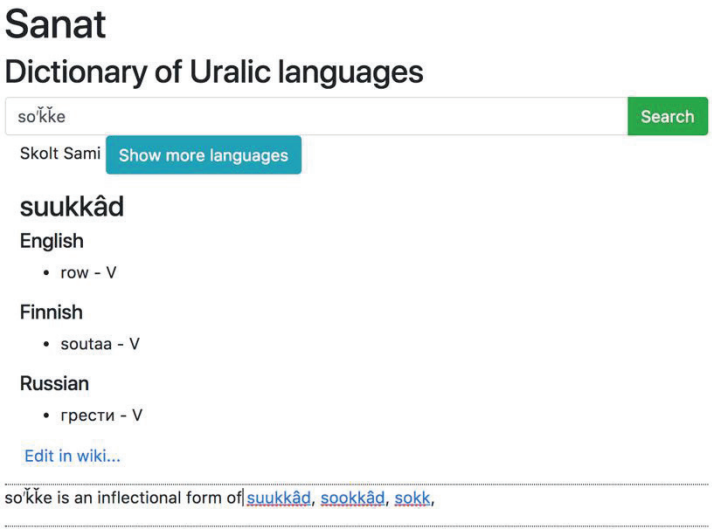

Figure 4: Easy search interface

The simplified search interface is depicted in Figure 4. It provides the functionality of picking the dictionary in which the words are searched, such as the Skolt Sami dictionary. Due to the highly inflectional nature of Uralic languages, a language learner might come across with a non-lemmatized form of a word. For this reason, our search interface incorporates morphological analyzers to lemmatize the user input word form. As seen in Figure 4, the search term used was *so'ǩǩe*, and the system found that it is an inflectional form of *suukkâd*, *sookkâd* and *sokk*. The inclusion of this feature is also motivated by previous research (Bergenholtz & Johnsen 2005) pointing out that online dictionary users use non-lemmatized word forms (the passive and imperative forms of a verb in their study) when consulting a dictionary.

It is also possible to use the same search to find words in the translations. This means that by inputting the English word *row*, the system will find the Skolt Sami entry *suukkâd.* The simplified search interface also provides a link to the full MediaWiki entry.

4.4 Semantic MediaWiki

Semantic MediaWiki (Völkel et al. 2006) is an extension that has been used in the past in the Language Bank of Finland MediaWiki environment with good experiences in the context of online dictionaries (Laxström & Kanner 2015). The extension makes it possible to link MediaWiki articles together based on shared semantic characteristics. The aim of the extension is to make semantic knowledge in a MediaWiki environment machine readable.

We use Semantic MediaWiki to gain access to a more graph-like representation of the dictionary data. We use it to enhance the MediaWiki entries with property tags in an automated fashion. The property tags are added or updated to the MediaWiki articles automatically always when new edits are made.

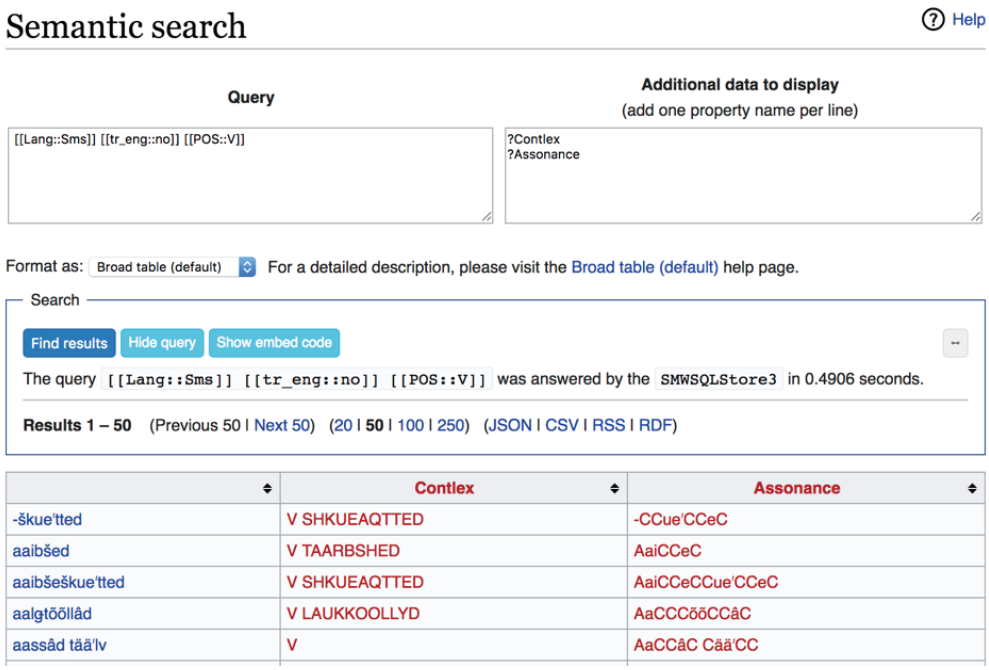

Figure 5: Semantic MediaWiki search

The property tags such as *tr_eng* or *Contlex* make it possible to query the dictionary information more effectively through the Semantic MediaWiki query interface. In Figure 5, we see how we can get a list of all Skolt Sami (*Lang::Sms*) words that do not have an English translation (*tr_eng::no*) and are

verbs (*POS::V*). We can also specify the property values we want to be visualized in the search results such as continuation lexicon (*Contlex*) and the assonance rhyme structure of each word (*Assonance*). These queries can be made within one dictionary or across multiple dictionaries stored in the system by altering the *Lang::* query parameter.

Furthermore, the extension allows us to access other entries of the same dictionary or entries of completely different dictionaries in the same system. This is achieved with the *pages that link here* functionality. This means that we can see, for each entry in the dictionary, if there is another entry possibly even in a different dictionary making a reference to a specific entry. Currently, these references might be translations, derivations or etymologies. In other words, just by having an etymological relation defined in the Skolt Sami dictionary, we can see the reference in the Erzya dictionary, for instance.

4.5 The API

As the dictionary uses morphological tools for different tasks, such as producing inflection paradigms when viewing an article in MediaWiki or lemmatizing input words in the simplified search view, the dictionary system has in built functionality that can be of a general interest when doing NLP for Uralic languages. This is the reason why we have decided to serve the morphological tools over an API that is currently usable through a Python library called Uralic NLP[5] (Hämäläinen 2018).

The underlying functionality relies on finite-state transducers based on the HFST tool (Lindén et al. 2013). These are openly available in the Giellatekno infrastructure (Trosterud; Moshagen & Pirinen 2013) in a source code format. Our API provides easy access to precompiled versions of the FSTs for morphological analysis, generation and lemmatization. In addition to the FSTs, the API makes it possible to get full JSON entries for words in the dictionary.

Apart from our own extended API, the standard MediaWiki API and Semantic MediaWiki API are available for the users. These provide a standardized access to the data stored in the MediaWiki side of the system, such as using the Semantic MediaWiki query language.

5 Lexicographical Difference of the XMLs and MediaWiki

Each dictionary is tailored to a different audience or user group. Whereas the XML dictionaries have been set up to act as virtually stand-alone databases that can be used for deriving any variety of output sets, the MediaWiki dictionaries have been set up to provide a less cluttered experience. In fact, the visible code in the MediaWiki presentation is less than what can be found in the XMLs. This design decision was taken to better support the end user goals when using the dictionary. A typical dictionary user is more likely to be interested in definitions and translations than metadata or FST specific information needed to produce the morphological analyzers. Visualizing too much information that is irrelevant for the user goals makes it harder for the user to find the relevant pieces of information. This would cause higher cognitive load which would take up more working memory (Paas, Renkl & Sweller 2003), which is the very thing we want to avoid with our design choice. Previously it has also been reported that extremely extensive entries cause difficulties in using the dictionary (Selva & Verlinde 2002).

The MediaWiki dictionaries utilize three different types of links. Etymon and audio links provide access to sites of external institutions, such as the Sami-language etymological database Álgu at the Institute for the Languages of Finland in Helsinki, and the Max Planck Institute audio archives in

5 Instructions and installation on https://github.com/mikahama/uralicNLP

Nijmegen. Cognate links facilitate navigation between languages in the namespace of our project on the CSC/Language Bank server, while compounding and derivation links enhance the navigation experience between compound words and their constituents in the same manner as derived words point to their derivational stems and morphemes. This interlinking provides a new alignment of semantic and morphological data not immediately accessible from the XML databases.

Not all homography is dealt with by means of Roman numeral identification. In fact, the development of XML dictionaries has led to the separation of homographs according to part-of-speech designation. When the MediaWiki dictionaries return all homographs to adjacent micro-entries within the macro-entries, micro-entries with the same part-of-speech designation are distinguished, as in the XML dictionaries, according to homograph enumeration, while other instances of homography are simply addressed with the help of part-of-speech marking.

Semantic tag values with synset distinctions are used in some language development at Giellatekno. In anticipation of shared meaning groups in source-to-multi-target-language dictionaries, this initial semantic tagging has been introduced in the XML dictionaries, where they reflect the same semantic tagging used in Constraint-Grammar disambiguation applied in the Giellatekno and Apertium infrastructures, and the ICALL infrastructure at Giellatekno. Initial outlines have also been drafted for editing semantic links that will enhance searches for various degrees of synonymy.

6 Discussion and Future Work

Our system is under continuous development, but it has reached a functional state. At the moment, we have several authors editing the Skolt Sami and Erzya dictionaries in the MediaWiki environment, while part of the dictionary editing is still ongoing in the XMLs. In this case of a handful of editors, the system has proved functional. The biggest limitation in the system, however, is the Semantic MediaWiki extension. Enabling the extension has a huge impact on the speed of the system when updating the entries in the MediaWiki side. We are currently finding ways to overcome this limitation.

The development has focused mainly on the technical side of the environment. Since the system is meant to be used by people with no linguistic or technical background, more research is needed in terms of usability and user experience of the system. This is especially needed and, in general, understudied in the context of editing the dictionary entries.

Giellatekno XMLs have the problem that they are not standardized by any means. This could be solved by remodeling the XML structure in a standardized TEI format. Since our system is built with multiple XML formalisms in mind, introducing a new TEI based format should not be too big of an issue. In fact, by writing a new template we can already start producing a TEI formatted version of the XML data.

The non-functional requirements of the system, reliability and scalability were solved by building the system on industry-scale open source technologies. These are MediaWiki, Django and MongoDB. Although these individual components are known to work reliably and scale well, there is a future problem of maintainability. This rises from the concern of the compatibility of our system with the future versions of MediaWiki and Django. Even during the two years we have been developing the system, a critical part of the MediaWiki API has already changed once. This required updates to our code in order to make our system work with the latest version of MediaWiki. This maintainability issue is solved by releasing the entire system as open source.

Currently, other users of the shared MediaWiki platform maintained by the Language Bank of Finland are showing interest in our system. Not only because it provides an already implemented way of

pushing dictionary data from another format to the MediaWiki system, but also because our system makes it possible to transfer the data edited in MediaWiki back to the original format.

7 Conclusion

In this paper, we have described our online dictionary system[6] with the aim of making XML based dictionaries editable by multiple users. We have described the advantages and limitations of Sematic MediaWiki in enhancing access to the dictionary data. Furthermore, the advantages of MediaWiki have been described. Our system is currently in use and has been proved to solve the problems we were set to solve with a small number of editors.

The dictionary system was originally developed for Skolt Sami, but we have successfully expanded it to cover 10 additional languages with minimal modifications. This has been possible due to the modular nature and ideology of separation of concerns embraced in the design process.

In addition to solving a dictionary editing problem, our efforts have made the XML formatted dictionaries available to a wider audience in an open MediaWiki format. The availability of these lexical resources online has a direct impact on the speakers and learners of these minority languages. The data has also been made available for research and technical purposes through the API of the system.

References

Álgu-tietokanta. (2002). Retrieved March 2018, from Kotimaisten kielten keskus: http://kaino.kotus.fi/algu/

Antonsen, L., Johnson, R., Trosterud, T. & Uibo, H. (2014). Generating Modular Grammar Exercises with Finite-State Transducers. *Proceedings of the second workshop on NLP for computer-assisted language learning at NODALIDA 2013*, (pp. 27-38).

Bergenholtz, H. & Johnsen, M. (2005). Log Files as a Tool for Improving Internet Dictionaries. *HERMES-Journal of Language and Communication in Business, 34*, 117-141.

Boicea, A., Radulescu, F. & Agapin, L. I. (2012). MongoDB vs Oracle - database comparison. *Third International Conference on Emerging Intelligent Data and Web Technologies* (pp. 330-335). IEEE.

Bothma, T. J., Gouws, R. H. & Prinsloo, D. J. (2017). The Role of E-lexicography in the Confirmation of Lexicography as an Independent and Multidisciplinary Field. *Proceedings of the XVII EURALEX International Congress*, (pp. 109-116).

Hämäläinen, M. (2018, January). UralicNLP (Version v1.0). *Zenodo*. http://doi.org/10.5281/zenodo.1143638.

Johannsson, E. T. & Battista, S. (2017). Editting and presenting complex source material in an online dictionary: the Case of ONP. *Proceedings of the XVII EURALEX International Congress*, 117-128.

Laxström, N. & Kanner, A. (2015). Multilingual Semantic MediaWiki for Finno-Ugric dictionaries. *Septentrio Conference Series*, *2*, pp. 75-86.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T. & Silfverberg, M. (2013). HFST — A System for Creating NLP Tool. *International Workshop on Systems and Frameworks for Computational Morphology*, (pp. 53-71).

Männamaa, K. & Iva, S. (2015). Võro-eesti-võro võrgosõnaraamat: synaq.org. In M. Velsker, & T. Iva, *Tartu Ülikoolo Lõuna-Eesti keele- ja kultuuriuuringute keskuse aastraamat* (p. 147−150). Tartu: Tartu Ülikooli Kirjastus.

Nicola, M. & John, J. (2003). XML Parsing: A Threat to Database Performance. *Proceedings of the twelfth international conference on Information and knowledge management* (pp. 175-178). ACM.

Paas, F., Renkl, A. & Sweller, J. (2003). Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist, 38*(1), 1-4.

6 The system has been released as open source in https://bitbucket.org/mikahama/saame/

Rueter, J. & Hämäläinen, M. (2017). Synchronized Mediawiki Based Analyzer Dictionary Development. *The Third International Workshop on Computational Linguistics for Uralic Languages*, (pp. 1-7).

Selva, T. & Verlinde, S. (2002). L'utilisation d'un dictionnaire electronique: une etude de cas. *Proceedings of the tenth EURALEX International Congress*, (pp. 773-781).

Töpel, A. (2014). Review of research into the Use of Electronic Dictionaries. In C. Müller-Spitzer, *Using online dictionaries* (pp. 13-54). Berlin - New York: De Gruyter.

Tender, T., Kallas, J., Laansalu, T., Nurk, T., Mihkla, M., Päll, P., Langemets, M., Soon, T. & Oro, K. (2017). *Eesti Keele Instituudi osakondade aruanded 201 7.* Tallinn: Eesti Keele Instituut.

Trosterud, T., Moshagen, S. & Pirinen, T. (2013). Building an open-source development infrastructure for language technology projects. *NEALT Proceedings Series*, *16*, pp. 343-352.

Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H. & Studer, R. (2006). Semantic Wikipedia. *Proceedings of the 15th international conference on World Wide Web (WWW '06)* (pp. 585-594). ACM.

Viikberg, J. (2008). Eesti keele kogud. In E. Parmasto, & J. Viikberg, *Eesti humanitaar- ja loodusteaduslikud kogud, seisund, kasutamine, andmebaasid* (pp. 95-112). Tartu: Tartu Ülikooli Kirjastus.