Electronic resources for Livonian

Vals Ernstīte
University of Latvia
Livonian Institute
Kronvalda 4-200, Riga, LV-1010, Latvia
vals.ernstreits@lu.lv

Abstract
Livonian is a Finnic language indigenous to Latvia. Presently, Livonian, which is listed in UNESCO’s Atlas of the World’s Languages in Danger as critically endangered, is spoken fluently by just over 20 people. Despite its regional importance, Livonian remains underresearched in many areas and also has a limited number of available resources. The core of the Livonian linguistic tools is formed by three databases that are entirely online-based and completely interconnected. The lexicographic database holds data of the Livonian-Estonian-Latvian dictionary (2012) and serves as the backbone for the morphology database and corpus; lemmas are also being added instantly from the corpus during the indexing process. The morphology database contains semi-automatically generated template forms for every declinable word found in the lexicographical database. This database is used for corpus indexing purposes, and – after indexation – for collecting morphological data from the corpus in order to statistically verify or point out differences in declination principles. The Corpus of Written Livonian contains a variety of indexed and unindexed Livonian texts and serves as a base for obtaining new lemmas for the dictionary as well as forms for the morphology database via the indexing process. The corpus has a dual purpose – it serves also as a repository of written texts in Livonian.

Taking into account the experiences and results acquired during the creation of linguistic resources for Livonian, the overall conclusion that can be drawn is that when available resources are minimal, solutions may be hidden in increasing workflow efficiency and data management in a way that allows one to extract maximum data with minimal effort. And also that sometimes simple manual on semi-automatic approaches may, in the long run, appear more efficient than fully automated solutions that are also more affected by ever-changing technologies.

Kokkuvõte
Liivi keel on UNESCO Maailma keelel oksitselt ohustatud keelena kantud Läti põlisekeel, mida tänapäeval räägib umbes 20 inimest. Vaatamata regionaalsele tähtsusele on mitmed liivi keele aspektid endiselt piisavalt uurimata ning ressurssid on piiratud.

Liivi keelele sisetsete ressursid tekkivad kolm omavahel integreeritud andmebaasi. (1) Leksika andmebaas sisaldab Liivi–Eesti–Läti sõnaraamat (2012) andmestikku ning on morfoloogia andmebaasi ja korpusse selgrooks; märgendamise käigus lisatakse sellele korpusest uusi lemmisid. (2) Morfoloogia andmebaas sisaldab osalt automaatselt genereeritud sabloonvormi iga leksiika andmebaasi leiduva muutsõna jaoks. Andmebaas kasutatakse korpusse märgendamisel vormide allikana ning märgendamise järel – morfoloogilise andmetikku kogumiseks korpusest olemasolevate vormide statistiliseks kirjutamiseks või erinevustele viitamiseks. (3) Liivi kirjakeele korpus sisaldab märgendatud ja märgendamata liivikeelseid teksti ning on uute märgendamise käigus saadud lemmade ja morfoloogiliste vormide allikaks. Korput kasutatakse ka liivikeelseid tekstele kogum.

Liivi keelele sisetsete ressursid loomise käigus saadud kogemused ja tulemusid viitavad sellele, et minimaalsete ressurside puhul võib lahendus peituda tööjärjekordisi ja tööjärjekordisi tähtsuse tõstmisega ühes andmetöötlemise nii, et minimaalsete vahenditega oleks võimalik hankida võimalikult palju erinevaid andmeid. Samuti võib lihtselt kasitseb toetav poolautomaatne lähenemine pikas perspektiivis osutuda efektiivsemaks kui täismootorsad lahendused, mida rohkem mõjutab ka tehnoloogia pidev muutumine.

1 Introduction
Livonian is a Finnic language indigenous to Latvia. During the 12th century Livonian was spoken across vast territories in Latvia along the Gulf of Riga, including the location of Latvia’s present-day capital. Livonians have contributed greatly to the historical development of the Baltic region and over time have shaped various layers of modern-day Latvian language and culture. Livonian is currently listed in Latvia’s language law (1999) as an indigenous language. The Livonian cultural space, including the Livonian language as its main component, has also been added to Latvia’s list of intangible cultural heritage (2018), beginning the journey towards the inclusion of Livonian into the corresponding UNESCO global list.

---

1 This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: http://creativecommons.org/licenses/by/4.0/

Proceedings of the fifth Workshop on Computational Linguistics for Uralic Languages, pages 184–191, Tartu, Estonia, January 7 - January 8, 2019.

184
Presently, Livonian, which is listed in UNESCO’s Atlas of the World's Languages in Danger as critically endangered, is spoken fluently by just over 20 people. The number of Livonians according to the most recent Latvian national census is 250, but the influence and importance of Livonian language and culture, however, reaches far beyond the numbers of Livonian speakers, as it is an important research subject not only for those interested in the Uralic nations, but also for scholars researching various aspects of Latvian as well as Uralic and Indo-European contacts.

Currently, despite its regional importance, Livonian remains underresearched in many areas and also has a limited number of available language resources. Livonian also suffers from extremely limited human resources – in terms of competent scholars and people possessing Livonian language skills and the fact that, due to the complex historical reasons, Livonian collections, scholars, and the Livonian community itself are either scattered across Latvia or exist abroad. In these circumstances a modern set of various linguistic tools is a necessity for solving numerous problems associated with research of Livonian, language acquisition, and accessibility of language sources, in order to ensure the competitiveness and continued development of Livonian.

2 Background

The first primitive linguistic database – a source for the later published first Livonian-Latvian dictionary (Ernstreit 1999) – was created 20 years ago as a FileMaker database file. However, at that time Livonian faced far more basic challenges associated with the arrival of the digital age such as creating fonts with Livonian letters, overcoming fixed sorting strings to match the needs of the Livonian alphabet, getting keyboard drivers for different platforms, etc.

Although several minor digital collections had already existed earlier, serious work on creating linguistic databases started as late as 2012 following the publication of the Livonian-Estonian-Latvian dictionary (LELD; ca. 13 000 lemmas). During the following year, the dictionary was transformed from its original text format into a database and published online (E-LELD). Following that, in 2015, the indexing tool Liivike was created, which used E-LELD as a lemma reference source and enabled the creation of a corpus of Livonian texts in phonetic transcription within the Archive of Estonian Dialects and Kindred Languages at the University of Tartu (murre.ut.ee). The database used for E-LELD and the tables of morphological patterns published in that dictionary were also used in the University of Helsinki project “Morphological Parsers for Minority Finno-Ugrian Languages” (2013–2014).

All of the aforementioned linguistic instruments, however, had their problems, e.g., the web version of the dictionary was created as a static database and therefore was complicated to update and correct. The dialect corpus utilized Uralic phonetic transcription and so was suitable only for research purposes (rather than, for example, language acquisition), its indexing system also allowed only for fully indexed texts to be uploaded or edited. This lead in many cases to “forced indexation”, especially for unclear cases, and indexation errors sometimes due to the poor Livonian language skills of the people doing the indexing. Also, due to the structure of the workflow, later corrections of various inadequate indexations were extremely complicated to correct, e.g., systematic indexation mistakes could be corrected in isolated textual units, but not across the entire corpus, etc. The morphological analyzer and
other tools created by Giellatekno within the University of Helsinki project worked nicely, but were made using an existing set of morphological rules. As further developments have clearly shown, morphological rules for Livonian remain at a hypothetical stage in many cases, as they still need to be further clarified and/or adjusted based on information gained from the corpus. However, the most severe flaw of all these previously existing systems and linguistic tools was the fact that they used the same initial source (the LELD database), but were also isolated, not providing any feedback with updates or corrections, requiring all efforts for keeping databases updated to be fully manual, thus being quite ineffective and never consistently performed. As a result, the understanding that a new approach for linguistic tools is needed gradually began to be form.

An effort to create a new set of linguistic tools for Livonian started out as a different project – when the Estonian-Latvian Dictionary (ELD; ca, 52 000 lemmas) had to be compiled from scratch and published as a joint effort of the Latvian Language Agency and the Estonian Language Institute within a timeframe of two and a half years. Since existing compiling tools proved to be too slow and incapable of handling such a large and complicated task within such a narrow time limit, a new online compiling system was developed to meet the project’s compiling needs. This system turned out to be extremely fast and productive, offering convenient opportunities for building dictionaries, shaping and reviewing word entries, executing various control tools (inverse dictionaries, compound controls, etc.), which ensured completion of this project on time.

In 2016–2017, this new system was developed further to suit the structure and needs of LELD (multilingual options, Livonian-specific sorting orders, source references, etc.), developing the online digital database of the Livonian lexicon (LLDB) based on LELD. It was then followed in 2017 by the online morphology database (LMDB) and the online corpus of Written Livonian (CWL) – these three databases form the core of the Livonian linguistic tools.

These databases are currently accessible for linguistic research purposes through registered-access modules. Their public part – mainly targeted towards language acquisition – is currently fully accessible in a separate section of the Livonian culture and language web portal Livones.net (lingua.livones.net). This section contains the online Livonian-Estonian-Latvian dictionary and word forms for each declinable word. In the future, it is planned to be supplemented with corpus data, a grammar handbook, and other linguistic tools to be developed on an ongoing basis.

3 Structure and work principles

All three databases forming the core of the Livonian linguistic tools – the lexicographic database, morphology database, and corpus of written Livonian – are entirely online-based and completely interconnected.

The general working principles within all the databases are based on simplified approaches – all necessary work is performed mainly by dragging, clicking, entering search criteria, or completing necessary fields. Workflow is made intuitive and no programming skills whatsoever are required by personnel involved in any of the processes. User controls are eased with visual attribution (e.g., color-indexed statuses, book-ready lemma articles, etc.).
The lexicon databases and corpus also include multilingual options – the possibility of adding translations of items into several languages (currently – Estonian and Latvian, but the addition of more languages is possible), in order to provide better use of materials by users with no Livonian skills.

The lexicographic database primarily holds updated data of the Livonian-Estonian-Latvian dictionary (LELD) and serves as the backbone for the morphology database and corpus; lemmas and variations are also being added instantly from the corpus during the indexing process and then developed further within the dictionary module.

The database contains lemmas and examples along with references to their sources, translations into several languages, basic grammatical data (word classes, declination, references). The system allows one to perform dynamic and creative changes within lemma articles (changing of meaning and example sequences, adding new meanings, cross-references, homonym identifiers, etc.). Book-ready lemma articles are displayed in this module dynamically during compiling in order to have an overview of the public presentation of data. In the compiling module, full grammar information is also displayed from the morphology database and in the near future it will be supplemented with a script that generates declinable word forms based on the indicated declination type.

The database also includes various statuses that allow one to identify the status of work performed (e.g., finalized, missing grammar, etc.) or to limit public access (e.g., technical lemmas from the corpus such as Latvian-like personal names or casual new borrowings). These may also be used for language standardization purposes.

This module also has several additional functions such as various search and selection options, a reverse dictionary, and also options for printing search results in the form of a pre-formatted dictionary.

The morphology database was initially built to ease work and the presentation of complex Livonian morphology, which contains a significant number of declination paradigms or types (currently 256 noun and 68 verb declination types).

The database contains full sets of paradigm templates (paradigm identifiers already served as grammar identifiers in the lexicographical database), example words and fields for all possible forms (separately for nouns and verbs). Based on these paradigm templates, semi-automated generation using simplified formulas (the initial form minus a number of letters to be deleted plus an ending according to the paradigm, e.g., Supine (lā’ mō) = Infinitive (lā’ dō) - 2 + mō) was used initially to generate template forms for every declinable word found in the lexicographical database connecting these forms with corresponding lemmas of that database. Since these formulas are connected with a corresponding paradigm they are also used to generate template forms for new lexemes belonging to that paradigm and acquired from corpora while indexing. No template forms were generated in case of significant stem changes (e.g., NSg őļaz > GSg allīd) – these were entered manually. The same also applies to rare forms, for which rules or endings are still unclear.

The result is accessible in matrix form offering an overview of all forms of words included in the corresponding paradigm and the automatic generation process also helped to reveal inconsistencies and subsequently to create new sub-paradigms. Also, based on this database, an overview of paradigm patterns is available for further methodological grouping. Template
forms of every declinable word can be moved to another – existing or new – paradigm retaining its current content, however content is fully editable as needed.

This database is used for corpus indexing purposes offering possible matches for indexation, and – after indexation – for collecting morphological data from the corpus in order to statistically verify word form templates or point out differences in declination principles. Although morphological paradigms linked to words in LELD and subsequently in the lexicography database have been collected over decades of field research, this statistical verification is done due to the fact that these paradigms still remain hypothetical to some extent, since there are many specific forms that are quite rare and may appear differently than initially assumed. This is the gap that feedback from the corpus can fill.

The database has already proven to be extremely useful and efficient for practical purposes in language acquisition. One of the key problems for people learning Livonian even after the publication of LELD has been that every time they needed to determine a particular form of a declinable word they had to use the paradigm number indicated in LELD and create the form themselves based on analogy with the corresponding example word in the morphological tables. This turned out to be very complicated and messy, especially for people with a Latvian background (including most of the Livonian community) not familiar with the “type word” approach used in Finnic languages. The morphology database allows one to abandon this entire process by providing a list of all necessary forms right in the lemma article by just clicking on the paradigm number.

The Corpus of Written Livonian contains a variety of indexed and unindexed Livonian texts and serves as a base for obtaining new lemmas for the dictionary as well as forms for the morphology database via the indexing process.

The corpus has a dual purpose – it serves as a linguistic source for research on Livonian, but also as a tool for researching other areas, e.g., folklore or ethnography. The corpus acts as a repository of written texts in Livonian. Sources used in the corpus are, therefore, quite varied. Although initially it primarily contained texts in literary Livonian (books, manuscripts, etc.), other written texts (folklore, texts in dialects, etc.) have been gradually added. The corpus also contains lots of metadata about the added texts, including their origin, dialect (if applicable), compiler or author, historical background, and other references. This data may also be used for narrowing searches – e.g., texts from a particular village, author, etc.

When texts are uploaded, they are split into subsections (e.g., chapters), paragraphs, sentences, and separate words, and then joined back together when the text is presented as whole. Previously uploaded texts are normalized so that they are represented using the unified contemporary Livonian orthography. Normalization mostly affects only orthographical representation, leaving, e.g., dialectal peculiarities intact. The same applies to texts written in phonetic transcription since there is no point in retaining phonetic details. This is, first of all, due to the fact that the purpose of the corpus is not phonetic research, second, that the Livonian contemporary orthography provides sufficiently detailed information on pronunciation, and, third and most importantly, due to the fact, that in the case of Livonian, instead of phonetic transcription one can speak of a phonetic orthography that displays texts according to certain rules of its own and not the actual pronunciation of those texts. This has also been confirmed by various later research projects revealing many important phonetic features found in Livonian, which are not reflected by texts written in phonetic transcription (e.g., length of various vowels, etc.).
During the indexation process a mandatory reference is made to the lemma and its form. In case of new lemmas or deviations from prior indexation, new records are generated in the lexical database and subsequently in the morphology database directly from the indexation module, using the default lemma form, reference to the form, and its source. Indexation itself is performed by selecting lexemes and their forms, and the lemma article view from the dictionary is available for the purposes of checking every form selected. For every word to be indexed, possible versions are offered based on either previous corpora statistics or the morphology database, and in most cases indexation can be performed by simply clicking to accept the offered combination or choosing a form from the list offered. It is also possible to search for a lexeme in the lexicographic database on the spot, choose a different, unlisted morphological form, or add a completely new lexeme. Indexed words and sentences are marked with color indicators in order to distinguish fully indexed, partially indexed, and unindexed parts.

All texts are available for searching as soon as they are uploaded and do not require to be fully or even partially indexed. While indexing, it is possible to leave an indexed word completely unindexed or marked as questionable, which does not limit the availability of texts for research. Since indexing languages with unclear grammatical rules involves lot of interpretation, it is also possible to add a completely independent second indexing interpretation (e.g., pinkōks ‘with a dog’: substantive, singular, instrumental ~ substantive, singular, comitative) or a reference to a completely different lemma and form (kōrandōl ’in the yard’: adverb ~ substantive, singular, allative).

It is possible to edit every sentence separately in order to eliminate possible mistakes in the original text, to add translations in several languages, and to set limitations for sentences, text parts, or entire texts with regard to public use for language standardization purposes. At every stage it is also possible to index texts or their parts, or to make corrections in existing indexations on the spot. This option is also available dynamically when entering the corpus from the search module.

4 Overcoming problems

Extremely small linguistic communities like Livonian are in quite a different position compared to larger language groups with more resources. Many linguistic instruments which seem entirely obvious to larger, or even not so large, language communities, simply do not work for much smaller languages, due to extremely limited resources – both in terms of people, and in terms of available legal, financial, and technical support. Also sometimes such communities face an entirely different set of problems to solve – problems, that at times are difficult to completely see and understand unless one is involved with such a community.

But there is also a bright side to this. Looking for solutions in unconventional cases may lead to unconventional approaches, which in the long run may appear more appropriate for the current situation. Below is a list of some of the problems addressed while building electronic resources for Livonian and solutions that may be of use also for other small linguistic communities.

The first of these is the fact that since the 1950s the Livonian community has been scattered across Latvia and also abroad. Likewise, Livonian researchers and resources also have traditionally been located across different institutions in various countries. This means that in
creating and using any Livonian resources, people from very different backgrounds are involved working from different platforms and different locations across the globe. So the only obvious solution to suit all of their needs is to create completely and purely online-based resources, which would be consistent, simultaneously accessible wherever they are needed, and function technically in the same way regardless of local technical solutions (e.g., fonts, operating systems, programs, etc.). Also, this solution would allow people with different linguistic or educational backgrounds to be simultaneously involved in the same processes, complementing each other’s efforts.

Secondly, when working with Livonian and presumably also other small languages, manual work is inevitable and only some processes can be fully entrusted to automated solutions, at least in their initial phases. For example, due to the existence of few and limited data, many automated features that are so common for larger linguistic communities cannot be applied – automated text recognition would not be effective since most of the texts are handwritten or printed at a poor level of quality. There is also considerable variation in orthographies. Automated indexing does not work because of a lack of clear and verified grammar rules and limited data. Machine translation cannot be executed properly due to a lack of those same grammar rules and limited data, etc.

Thus, when developing linguistic tools for small linguistic communities, the main focus should be on helping to maximize the efficiency of all areas of manual work, supporting semi-automated solutions instead of fully automated approaches, which – due to insufficient or occasionally incorrect input data – may lead in the long run to completely messing up the entire effort by, e.g., creating a large number of misinterpretations. Also, since linguistic sources for small languages are significantly smaller anyway, the creation of fully automated solutions may also be questionable from the perspective of the effort necessary to create them versus the actual benefits gained from their creation.

Thirdly, there is a disadvantage of limited sources that in the case of Livonian has actually been exploited as an advantage. For smaller linguistic communities there is the possibility to connect different language resources. In the case of more widely-used languages, such resources are usually developed by separate institutions. Smaller languages can unite such resources under one roof, interconnect them, make one resource supplement another without any great additional effort, and ensure overall data consistency, thus supplementing lack of quantity with quality.

Databases created for Livonian, for example, also allow one to simultaneously perform linguistic research on the language while dynamically setting the language standard (e.g., adjusting morphological templates, suggesting better lexemes, omitting from public view poor quality texts, etc.). And, last but not least, since language materials also contain important cultural value, it is important to retain their availability as textual units for research and use that may have nothing to do with linguistics.

Such an approach allows one to extract maximum data from limited sources with minimal effort. In a sense it is reminiscent of Livonian Rabbit Soup, which has nothing to do with rabbits and is made as an extra dish by simply not throwing out the water left over from boiling potatoes for dinner.

The fourth problem is a lack of personnel with sufficient linguistic and language skills. This is one of the most serious issues that is faced by any smaller language. In the case of
databases created for Livonian, this problem has been addressed by two separate approaches. The first is to simplify work methods and technical solutions, bringing them down to simple familiar actions mostly performed using a computer mouse such as clicking, choosing from drop-down menus, dragging, etc., which also helps to contain possible mistakes. Secondly, and most importantly it is addressed by overall principles of database performance and workflow.

This means that people with lesser skills only perform actions matching their skill level. For example, they transcribe texts from manuscripts following a set of normalization rules, but final normalization prior to adding the texts to the database is performed by better-skilled scholars. This principle is also integrated into the corpus indexing principles where lesser skilled personnel only index simple items of which they are completely certain (such items also happen to make up most of the texts to be indexed), leaving complicated cases for more skilled personnel. Ultimately, this saves time and effort for everyone involved.

Closely connected to this is also fifth problem that is the finalization of database content. In most cases, content of databases is usually completely prepared and finalized before making it available for further use. However, in the case of Livonian and also perhaps in the case of many other small languages, such preparation and finalization of content is sometimes quite complicated. This is mainly due to a lack of sufficient people or time to perform the necessary work, but also due to unclear interpretations. Concerning mandatory need to finalize content in corpora, in many cases this leads to “forced indexation”, which is a significant source of misinterpretations and leads to a later necessity for additional work involving elimination of such incorrect indexations. Also waiting for completion and finalization of content may limit or significantly postpone its use for research purposes.

In the Livonian case, this is addressed by making all content available immediately, e.g., texts are fully searchable right after they are uploaded and there is no requirement for them to be indexed at all. During indexation it is also possible to index the whole text, index it partially, mark it as questionable, or add different interpretations. At the same time, all actions (indexation, adding lemmas, etc.) can be performed at any stage of working with the databases, even during research of some other subject. However, indicators are used for marking completed workflows (completed lemma articles, completely indexed sentences, etc.). This means that all resources are fully usable, each to a certain extent depending on readiness, of course, and unclear cases, at the same time, can be left unclear until they can be resolved at some point in the future or indexed purely as an interpretation leaving it for final attention at a later time.

Taking into account the experiences and results acquired during the creation of linguistic resources for Livonian up to this point, the overall conclusion that can be drawn is that when available resources are minimal, solutions may be hidden in increasing workflow efficiency and data management in a way that allows for maximum output with minimal effort. And also that sometimes simple manual on semi-automatic approaches may, in the long run, appear more efficient than fully automated solutions that are also more affected by everchanging technologies.

5 Future plans
In the near future, there are plans to continue development of databases within several projects in Latvia and Estonia. Upcoming projects include addition of a Livonian folktale
corpus, which enables the handling of various subdialects of Livonian; transfer of the existing phonetic transcription-based corpus (E–LELD) to lingua.livones.net; and construction of a separate topographical map-based Livonian place names database that would eventually allow for the development of a universal tool for areal research and mapping of linguistic patterns, using information from other already existing Livonian databases.

Acknowledgements
This study was supported by the Latvian Ministry of Education and Science State research program “Latvian language” (VPP-IZM-2018/2-0002).

References
CWL= Corpus of Written Livonian (2017). Rīgõ: Līvõ kultûr sidām. Available online at <http://lingua.livones.net/lv/module/korpuss>. Accessed on 13.11.2018.
ELD =Ernštreits, Valts; Marika Muzikante, Maima Grīnberga (2015). Eesti-läti sõnaraamat=Igaunu-latviešu vārdnīca. Tallinn: Eesti Keele Instituut.
E–LELD = Viitso, Tiit-Rein and Valts Ernštreits (2013). Līvõkīel-ēstikīel-leţķīel sõnārōntōz. Liivi-eesti-lāti sõnaraamat. Lībiešu-igaunu-latviešu vārdnīca (web version). Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra. Accessible online: http://www.murre.ut.ee/liivi/. Accessed on 13.11.2018.
Ernštreit, Valt (1999) Līvõkīel-leţķīel-līvõkīel sõnārōntōz=Lībiešu-latviešu-lībiešu vārdnīca. Rīga: Līvõ kultûr sidām.
LELD = Viitso, Tiit-Rein and Valts Ernštreits (2012). Līvõkīel-ēstikīel-leţķīel sõnārōntōz. Liivi-eesti-lāti sõnaraamat. Lībiešu-igaunu-latviešu vārdnīca. Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra.
LLDB= Viitso, Tiit-Rein and Valts Ernštreits (2016). Livonian Lexicographic Database. Rīgõ: Līvõ kultûr sidām. Available online at <http://lingua.livones.net/lv/module/vardnica/>. Accessed on 13.11.2018.
LMDB = Viitso, Tiit-Rein and Valts Ernštreits (2017). Livonian Morphology Database. Rīgõ: Līvõ kultûr sidām. Available online at <http://lingua.livones.net/lv/module/vardnica/list_vtipi>. Accessed on 13.11.2018.