Lexical Tools for Low-Resource Languages: A Livonian Case-Study

Valts Ernštreits

The University of Latvia Livonian Institute, Kronvalda 4-220, Riga LV1010, Latvia
E-mail: valts.ernstreits@lu.lv

Abstract

This article focuses on the empirical experience and conclusions, resulting from the creation of language research and acquisition tools for Livonian – one of the smallest languages in Europe.

A cluster was created for Livonian containing three interconnected databases, each with distinct types of data – lexical, morphological, and a corpus. The lexical database contains the lemmas and their data, the morphological database stores morphological forms, while all textual material, including the dictionary examples, is in the corpus. When indexing the corpus, every word refers to a lemma in the lexical database and its morphological information (new lemmas are added prior to indexation), ensuring consistency of the language data, and from each database the full data set of the other databases can be accessed.

The function of each cluster is to extract the maximum amount of information from limited data sources. While technologies designed for languages with a large number of speakers focus on using quantitative methods and automation to extract qualitative information from a large and constantly expanding amount of linguistic data, the main function of technologies designed for small languages is to extract the same type of information from a limited and largely static data set.

This article also examines a string of problems faced when working with a small amount of resources (inadequate language data, insufficient personnel, lack of rules for automating processes, etc.) and methods for resolving these problems in the case of Livonian.

Keywords: Livonian; low-resource languages; lexicography; corpora; data collection

1. Introduction

Livonian, one of the smallest languages in Europe, at present is spoken fluently by ~20 people[1]. Although currently listed in UNESCO's Atlas of the World's Languages in Danger as critically endangered (unesco.org), historically, the Livonians have had a

[1] According to the most recent census, 250 Livonians live in Latvia (csb.gov.lv); however, the majority of them do not speak Livonian and a reliable estimate of the number of speakers cannot be made. Due to the scattered nature of the Livonian population and the complex language situation, since the most recent Livonian speaker census in 1935–1937 (Blumberga, 2006), no other attempts have been made at assessing the total number of Livonian speakers. However, in Summer 2019, there are plans for a sociolinguistic pilot project to be carried out by UL Livonian Institute researchers to determine the number of Livonian speakers and their level of proficiency.

significant role in the development of modern-day Latvia and the entire Baltic Sea region. For this reason, Livonian language resources must be accessible not only to the Livonian community for language revitalization work, but also to society at large. Though the number of Livonian speakers is extremely small, Livonian requires the same opportunities and language tools as any other language. This article is focused on the experience and conclusions resulting from the design of technical language support tools for Livonian over the course of the last five years. Some aspects are also discussed in previous studies (e.g., Ernštreits 2019).

A seemingly small user base, associated limitations in being able to access financial resources as well as institutional lack of interest are not the only problems one encounters when creating modern language tools for exceptionally small languages like Livonian. The small amount of speakers also places a limit on the number of people who could potentially be involved in creating these language tools, which means that every potential language tool must be evaluated based on the actual possibilities for creating it and also its effectiveness. This same problem likewise affects the accessibility of the end product; for these tools to be usable by a wider audience for the purposes of language research and learning, one has to already consider the fact that these tools will need to be equipped with translations into one or more other languages (usually: Latvian, Estonian, English).

An added challenge faced by Livonian is that following the Second World War and the Soviet occupation, Livonian speakers were scattered across Latvia and the world. The same is true for Livonian language sources and researchers, which are located at various different institutions (Ernštreits, 2012). This means that when creating or using any resource for Livonian, people from very different backgrounds are involved and are working from different platforms and locations around the globe.

Another important aspect is that the grammars of small languages are often insufficiently studied. As a result, there are many processes which cannot be automated due to lack of knowledge of grammatical rules, while other alternatives, such as solutions based on neural networks, do not function well due to insufficient data. Additionally, sources of language data have been recorded at different times and so they are often written using different transcriptions[2], which limits the possibilities for people without existing specialized knowledge from using these sources and makes it difficult to process these texts electronically. However, the primary problem is that small languages consistently suffer from a lack of sufficient institutional interest, as well as inadequate data sources resulting from insufficient documentation.

The abundance of available resources is also the primary distinguishing factor when

[2] The problem faced by Livonian is the wide-ranging use of phonetic transcription, which, moreover, is not used for its basic function – accurately depicting pronunciation – but rather as a systematic means for writing down sources, including lexical sources (Ernštreits, 2011).

creating technologies for different languages. Technologies designed for languages with a large number of speakers focus primarily on using quantitative methods and automation to extract qualitative information from a large and continuously expanding amount of linguistic data. The primary function of technologies designed for small languages is to acquire that same information from a limited and largely static data set, primarily using qualitative methods in an effort to extract the maximum amount of information in circumstances where the available human resources and opportunities for automating this entire process are also limited.

A database cluster containing lexical morphological databases as well as a Livonian language corpus was created to resolve all of the aforementioned problems faced by Livonian. Its function is to ensure information acquisition from the limited Livonian language sources, while simultaneously optimizing the tasks carried out by the personnel working with the database and creating a base for further expanded use of both existing and future databases.

2. Creating the database clusters

2.1 Earlier Livonian language dictionaries and databases

The history of Livonian language dictionaries is relatively long. The first Livonian dictionary (Livonian-German-Livonian) was published in 1861 (~9,000 lemmas; SW); it was followed in 1938 by a Livonian-German dictionary (~13,000 lemmas; LW). Both of these publications were primarily intended for researchers, and the Livonian entries were written using phonetic transcription (Ernštreits, 2011).

The first Livonian dictionary (Livonian-Latvian-Livonian) intended for general use, and in which all Livonian entries were written using the orthography of the Livonian literary language, was only published in 1999 (~5,000 lemmas; LLLS 1999). This was also the first collection of Livonian vocabulary compiled using electronic tools. It was assembled, beginning in 1995, from entries in the 1938 dictionary using the Filemaker database software, though due to various reasons the primitive system used for compiling this dictionary was not further developed. However, for its time it was somewhat advanced. One of the first Livonian fonts and also lemma-sorting algorithms were designed for this dictionary, as well as the first Livonian keyboard drivers, the principles of which continue to be used up to the present day.

After a lengthy hiatus, in 2012, the most extensive lexicographic publication in the Livonian literary language – the Livonian-Estonian-Latvian Dictionary (13,000 lemmas; LELD) was published. The basis for this dictionary is the nearly 40 years of work by Estonian researcher Tiit-Rein Viitso, who collected and compiled Livonian vocabulary, language examples, and morphology. For understandable reasons, the basis of the dictionary was prepared using analogue methods. In the project's final phase, the information from the card index was transferred to MS Word format.

During the next year, the dictionary was transformed from its original text format into a database and published online (murre.ut.ee). Following that, in 2015, the indexing tool *Liivike* was created, which used this database 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. The aforementioned electronic dictionary 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 tools, however, had their problems, e.g., the web version of the dictionary was created as a static database, and therefore was difficult 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 sometimes indexation errors 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 difficult to correct, e.g., systematic indexation mistakes could be corrected in isolated textual units, but not across the entire corpus, etc.

The morphological analyser and other tools created by the University of Helsinki project worked well, but were made using an existing set of morphological rules and were therefore static and sometimes incorrect. 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 database based on LELD digital data), but were also isolated, not providing any feedback with updates or corrections, requiring all efforts to keep the databases updated to be fully manual, and thus being quite ineffective and never performing consistently. As a result, the understanding that a new approach to linguistic tools was needed gradually began to form.

2.2 The precursor of the Livonian language database cluster – the Estonian-Latvian dictionary

It could be said the events outlined above happenstance led to the creation of the cluster and its databases. In 2013, a working group was formed with professionals from Latvia and Estonia in order to compile both the print and electronic versions of the Estonian-Latvian and Latvian-Estonian dictionaries. These dictionaries, containing 40,000 lemmas each, had to be compiled from scratch and published as a joint effort of

the Latvian Language Agency (Estonian-Latvian; ELD) and the Estonian Language Institute (Latvian-Estonian; LED) within a timeframe of two and a half years.

Originally, the Estonian side was to use its own lexicographic working environment EELex for compiling the dictionaries; however, once testing began, the Latvian side concluded that this system was outdated and worked too slowly. For example, an average of 1.5 minutes was needed to open an entry, make an edit, and close the entry in this system, which meant that a compiler, who needed to compile at least 30 entries per day in order to meet the project deadline, would lose at least 45 minutes of work time per day. In addition, though this system could be used online, it was possible to use it with only one type of operating system and one type of browser.

When there were no results after almost half a year of attempts to resolve the issues with the EELex system relating to the speed of operation and other aspects of compiling entries, the Latvian side decided to begin immediate work on a solution. Within 48 hours they had constructed a temporary online system not connected with any particular operating system, which decreased the opening/closing time for each entry to two seconds and permitted the user to see the entry with its final formatting as information was added to it, to move examples and entire definitions between groups without difficulty, search for entries, view the completion status of each entry as well as use data from the Estonian Language Frequency Dictionary (EKSS), the unified corpus (cl.ut.ee), and the Glossary of Estonian Basic Vocabulary (EKPS) for selecting lemmas and also print out any part of the dictionary or print out the full dictionary in its final formatting. Unexpectedly, this system proved to be so productive that the decision was made to continue work on the Latvian side using this system. During the course of the project it was supplemented with a string of other tools necessary to ensure the quality of the final product – a reverse dictionary, compound word inspection, and other tools.

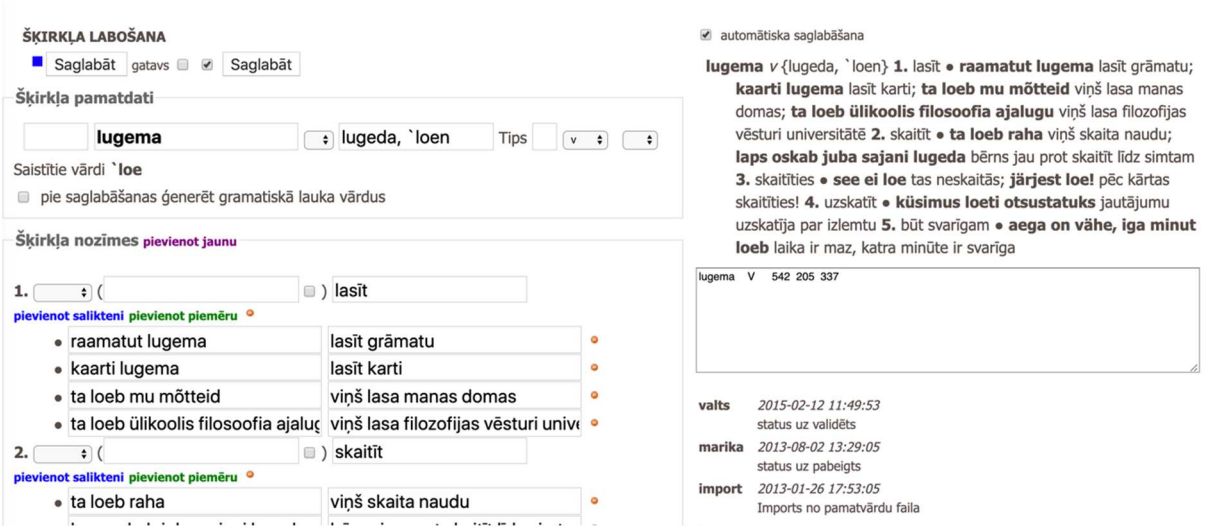

Figure 1: An inside view of the Estonian-Latvian dictionary compiling module.

This system, which was built using our own resources in conjunction with corpus data for lemma selection, proved to be one of the main steps in compiling the 1,096-page-long Estonian-Latvian dictionary. This work was done over an incredibly short time period without sacrificing the quality expected from lexicographic sources.

A logical question that might emerge from this is whether it is possible to speed up this work even more by utilizing parallel corpora to find correspondences. The Estonian side proceeded down exactly this experimental path. They worked in cooperation with the Latvian language resource company "Tilde" to create the basis for a dictionary compiled in an automated manner utilizing parallel corpora; however, this work did not take into account the aspects discussed in this article

Thus, in addition to the term "low-resource languages", an additional term should be used – "low-resource language combinations", i.e., those language combinations without parallel corpora or corpora formed from a limited number of sources, translators, documents, and text genres. This leads to the problems noted in the introduction, namely that in circumstances characterized by insufficient information (as is the case for the Estonian and Latvian language combination) automatic methods cannot be used due to inadequate data. The aforementioned experimental Latvian-English dictionary, which was essentially a structured word list collected from parallel corpora without any word use examples, was criticized for its low lexicographic quality (Bušs, 2015).

2.3 The formation of the Livonian language database cluster

Following the successful completion of the Estonian-Latvian dictionary project in 2015, the decision was made to adapt the lexicographic system created for compiling the Estonian-Latvian dictionary to the needs of the LELD. In 2017, it was supplemented with fields for correspondences in a second language, adjusted to be used with the Livonian writing system and Livonian alphabetic sorting, fields were added for the supplementary information found in the LELD database, but not included in the print version – the sources for the lemmas, correspondences, and examples – along with other necessary additions.

Following the beginning of work on the new dictionary and its publication online (livones.net), it was concluded that from the perspective of language acquisition it was still vital to resolve one of the most troublesome Livonian language problems faced by everyday users – the method for displaying the inflectional morphology of words in the dictionary.

Livonian morphology is relatively complicated. In order to show word inflection, Livonian follows the Estonian example of using word types (usually these are noted in each entry with a numeral following the lemma), which are a model used to show the changes that occur for all words within a particular word group. Livonian has 256

declension types and 68 conjugation types. It is impossible for a user to remember all of these, and it is also complicated to form the inflectional forms of other words by analogy. In order to simplify looking up forms and to free users from needing to constantly use a word type table, a morphological database was created, which utilized the templates included in the LELD and in a partially automated manner generated a template of declination or conjugation forms corresponding to each lemma. As a result, users can see all the forms of that word by clicking on the numeral corresponding to the word type of that word.

After the creation of the morphological database and the active use of the databases, subsequent research showed that the dictionary examples contain lemmas which are not found in the dictionary itself, as these examples had been used to illustrate the use of other lemmas. As a result, the idea arose to supplement both existing databases with a corpus, which would extract vocabulary from texts – using the examples in the LELD as its first text – and collect associated morphological data for the morphological database. This morphological data would be used to test the accuracy of the morphological form template and to gather information about the morphological forms not included in the templates. Since the first part of the cluster was the lexical database, it was logical to connect all subsequent databases to it. However, it was not possible to fully gauge the effectiveness of this solution until the third part of the cluster – the corpus – was completed.

Currently, these three databases are accessible for linguistic research purposes through registered-access modules (lingua.livones.net). Their public parts – mainly targeted towards language acquisition – are currently fully accessible in a separate section of the Livonian culture and language web portal Livones.

3. Cluster operating principles

The cluster is composed of three Livonian language databases – the lexical database, morphological database, and corpus – and consists of interconnected data archives, which have been compiled using the Livonian literary language and are completely editable and usable online. The lexical database forms the backbone of the cluster and each section contains a different type of data. All databases are built with a relational database structure and JSON objects, and the dictionary engine is powered with a PHP application for the backend and simple API calls on the frontend.

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., colour-indexed statuses, book-ready lemma articles, etc.).

3.1 Lexical database

The lexical database contains only information about the lemmas, parallel forms, semantics, representations in other languages, and the source of the lemmas. The only grammatical information it contains is the word class, word type, and a reference to the use of the word in singular and plural. The other grammatical information concerning the lemma – the word form template created by generating the forms using the word type as a model – are stored in the morphological database, with the lexical database only containing links to these forms.

All the example texts and the references to their source in the lexical database are shown as data from the corpus. The lexical database only contains a link to the respective sentence in the corpus. The lemma is also linked to every indexed use of the lemma in the corpus. The original examples used in the LELD were also transformed into a separate part of the corpus and their separation from the lexical database was one of the main changes undertaken in the process of connecting all of the parts into a single cluster. This was done to prevent duplication of data and ensure the consistency of the data across the entire cluster.

The lexical 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.

3.2 Morphological database

The morphological database contains fields for all known word forms and parallel forms (the set of forms depends on the word class). They are partially filled with word form templates, which are generated according to the word type example, a process which is partially automated with the help of simplified formulas. These formulas are also used in generating form template sets for new lemmas to be included in the lexical database. The database also contains empty fields for rarely encountered forms or those not included in the morphological examples found in the LELD, as well as those with formation principles that remain unclear and also parallel forms.

The morphological 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 verify word form templates statistically or point out differences in declination principles. Although morphological paradigms are linked to lemmas and 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 or may be statistically not dominant. This is the gap that feedback from the corpus can fill.

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 helps to reveal inconsistencies and subsequently to create new sub-paradigms. Moreover, based on this database, an overview of paradigm patterns is available for further methodological grouping. It is also possible to change a word's type within a word class with the same morphological principles without losing existing data which had already been generated or links to the lexical database or corpus.

3.3 The Corpus of Written Livonian

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 morphological 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, as it also simultaneously serves as a repository of written texts in Livonian. Sources used in the corpus are, therefore, quite varied. Although initially it mostly 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 entire text is presented. Previously uploaded texts are normalized so that they are represented using the unified contemporary Livonian orthography. Normalization mostly affects only orthographical representation, leaving things such dialectal peculiarities intact. The same applies to texts written in phonetic transcription.

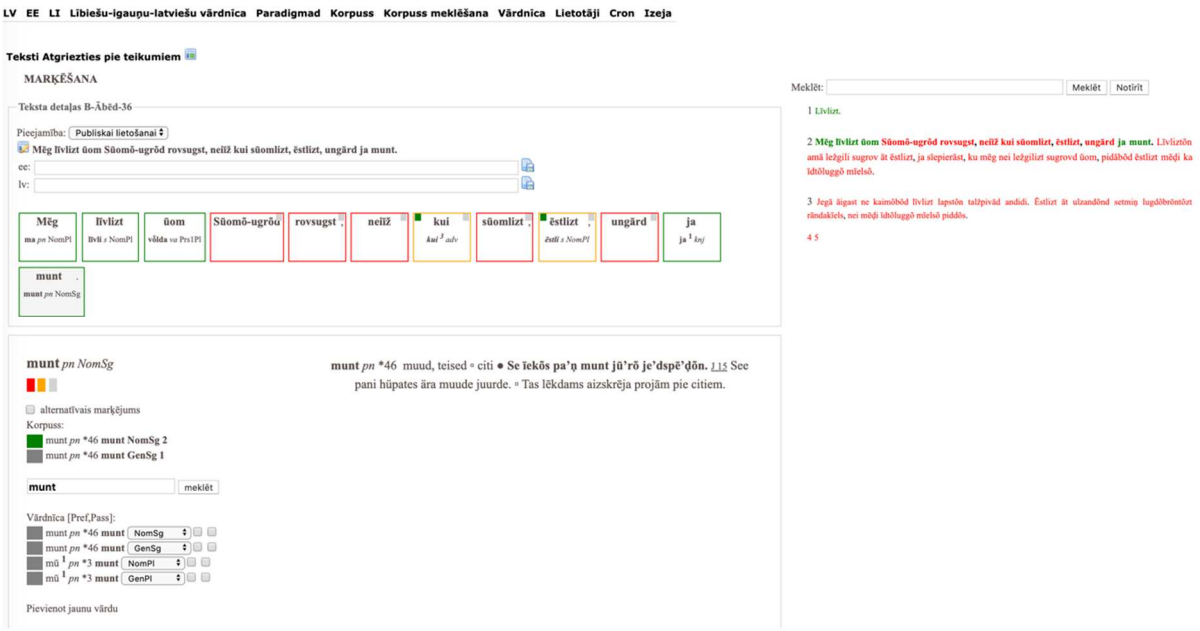

Figure 2: An inside view of the indexation module.

During the indexation process a mandatory reference is made to the lemma (lexical database) and its particular form (morphological database). In the case of new lemmas or deviations from prior indexation, new records are generated in the lexical database and subsequently in the morphological 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 morphological 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 lexical database on the spot, choose a different, unlisted morphological form, or add a completely new lexeme. Indexed words and sentences are marked with colour 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 have 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 a lot of interpretation, it is also possible to add a completely independent second indexing interpretation (e.g., *piņkõks* 'with a dog': noun, singular, instrumental ~ noun, singular, comitative) or a reference to a completely different lemma and form (*kōrandõl* 'in the yard': adverb ~ noun, singular, allative).

Indexation sources	
Corpora	Primary indexation source – **candidates** from previously indexed forms, statistics, sentence translations, etc.
Morphology database	Secondary indexation source – **candidates** from forms listed in the database
Lexical database	**Lemma information** (semantics, grammar information, etc.)
Indexation process	
General principles	The indexation process attributes a lemma and a form to every indexed word. Indexation is performed by clicking (all steps), picking from a drop-down menu (step 4), entering search criteria (step 5), correcting the lemma form (step 6). When clicking any candidate, the lemma article data is displayed. When indexation is completed, the module automatically jumps to the next word.
Step 1	Primary choice – click to accept the most popular indexation match from corpora statistics (the choice offered inside the yellow box at the sentence level) or indicate it as not to be indexed (e.g., number).
Step 2	If not, click to choose an alternative indexation match from the corpora statistics.
Step 3	If not, choose an alternative matching lemma and form from the morphology/lexical database
Step 4	If not, choose an alternative matching lemma from the morphology/lexical database and choose an alternative form from the drop-down menu.
Step 5	If the lemma is not found (e.g., a very different form), manually search for an alternative lemma within the lexical database, then return to step 4.
Step 6	If the lemma is still not found, add a new lemma (semi-manual, when adding a new lemma to the lexical database the word in the form as found in the sentence and the source reference are taken together), correct the lemma form and add the word category, then return to step 4.
Step 7	If necessary, add a second alternative indexation by clicking the checkbox and repeating steps 2 to 6.
Step 8	If in doubt about the indexation outcome, set the status to yellow (unfinished) or red (clear indexation).
Indexation output	
Corpora	Statistics and indexation candidates
Morphology database	Actual forms, statistics
Lexical database	References to the existing lemmas / new lemmas added; source references; example data – references to the sentences; semantics – references to the sentences and meanings in the lemma article (planned)
Other options	
Translations	Translations into several languages may be added.
Corrections	Corrections may be made in the sentence itself, in any indexed items, in translations, etc.
Sentence management	Public access may be restricted, if needed; a sentence can be added to one or multiple lemma articles as an example.

Table 1: Indexation scheme and options available in the indexation module.

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 portions, 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 to existing indexations on the spot. This option is also available dynamically when entering the corpus from the search module.

4. Problems and solutions

A cluster consisting of three interconnected databases has allowed Livonian to turn a lack of resources into an advantage. The general trend for large languages is that different institutions control and develop their own type of database – lexical and morphological databases, corpora, and tools for language acquisition. The benefits of sharing these resources remain untapped not only due to differing interests, but also often due to the incompatibility of these resources. The reasons for this are not always exclusively technical. At the same time, in the Livonian case, the lack of institutional resources has resulted in a solution which ultimately ensures the compatibility of various linguistic data, data consistency, avoiding data duplication, and provides high quality data processing. It also makes it possible to use various types of complementary data from a single resource for research as well as language acquisition, with the option of combining linguistic data with other types of data.

The Livonian experience shows manual work is inevitable that when developing linguistic tools for small linguistic communities, and only some processes can be fully entrusted to automated solutions, at least in their initial phases. For example, even automated text recognition would not be effective since most of the texts are handwritten or printed at a poor level of quality. Also, as only a small proportion of them are available electronically, automated indexing does not work because of a lack of clear and verified grammar rules, limited data, etc.

Thus, one of the main sources for improving efficiency can be found in maximizing 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 undermining the entire effort by, for example, creating a large number of misinterpretations. Also, since there are significantly fewer linguistic sources for small languages 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.

Increasing productivity is one of the main approaches for compensating for insufficient personnel with an adequate level of relevant linguistic and language knowledge. A considerable lack of human resources affects not only the Livonians, but nearly any small language. In the case of Livonian, this has been addressed using two different approaches.

The first approach is to simplify work methods and technical solutions, bringing them

down to the level of simple, familiar everyday actions such as clicking, choosing from drop-down menus, dragging, etc., which also helps to limit possible mistakes.

The second approach addresses the overall principles of database performance and workflow, which are organized so that personnel only complete the tasks for which they are objectively qualified. 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 more skilled scholars. This approach is also integrated into the corpus-indexing principles, where less-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.

Another means for increasing effectiveness is to ensure that when the system is being created, it is coded so as to allow many types of uses as well as dynamic and creative options for adapting the databases and their contents to serve different uses. The databases created for Livonian, for example, also allow one to simultaneously perform linguistic research on the language while dynamically setting the language standard, which is relevant for many insufficiently studied and standardized languages (e.g., adjusting morphological templates, suggesting better vocabulary, excluding poor quality texts from public view, etc.). In addition, language materials can also hold significant cultural value, so it is possible to keep them available as textual units for research and other uses unrelated to linguistics.

Incidentally, in the Livonian case something that has been important and seems elementary by current standards is that technical independence is built into the foundations of this system. This relates not only to being able to access all functions online, which makes it possible to work with the data in the database and expand the database regardless of one's actual physical location, but also that it functions independent of any operating system, browser, or other programs. Likewise, the user does not need to have language support (fonts, keyboard drivers) for Livonian or any other language used in the system, which in the past had turned out to be a significant barrier to, for example, Livonian language acquisition.

Surprisingly, one of the most important factors in developing and using the language database cluster for Livonian has turned out to be that it is left unfinalized. In most cases, the content of databases is usually completely prepared and finalized before making it available for further use. However, such finalization of content tends to be quite complicated, due to a lack of sufficient people or time to perform the necessary work, though mostly due to inadequate knowledge of clear rules and relevant studies. With regard to a mandatory requirement that corpora content be finalized, in many cases this leads to "forced indexation", which is a significant source of misinterpretations and leads to additional work later on involving the elimination of

incorrect indexations. Moreover, waiting for completion and finalization of content – e.g., lemma articles, morphological standards, etc. – may limit or significantly postpone its use for research or language acquisition.

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 entire 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 while researching some other subject (though indicators are used for marking completed workflows).

This means that all resources are fully usable, each to a certain extent depending on readiness, of course, and at the same time unclear cases can be left unclear until they can be resolved at a future point or indexed purely as an interpretation, leaving them for final attention at a later time.

Leaving the database unfinalized also prevents the work from stalling – for example, if it is not possible to precisely define a word or place it in a specific morphological category. This makes it possible to work with other, achievable tasks, as there is always much to be done when it comes to working with small languages.

This also allows for the application of the open-contribution principle, where every researcher using these databases is able to contribute little by little in the areas on which they are working within a particular study, by adding indexations or corrections, resolving unclear cases, contributing translations, etc.

The combination of all of these efforts makes it possible to extract the maximum amount of 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.

5. Future plans

Though initially this system was created as a Livonian language data archive and a tool for language research, standardization, and acquisition, its principles can also be adjusted to suit other types of studies by supplementing it with other digital archives (containing images, audio recordings, video, 3D scans, data from other databases) as well as other information. In this way, the synergy and coordination among various archives can create a rich, high-quality tool suitable for multi-faceted studies in other fields or for interdisciplinary research, for the effective use of data and research results for the preservation, maintenance, and development of any low-resource language and cultural community existing in circumstances characterized by limitations on data, personnel, financing, and other resources.

One of the projects the UL Livonian Institute will undertake in the near future that will further develop this platform is the creation of a Livonian place name database. This database will link Livonian place names found in the corpus texts with their corresponding Latvian place name cartographic and geospatial data. This will be followed by linking the existing databases geographically with their sources, using existing metadata in the corpus and lexical database relating to the language informants and the data-recording location. This will make it possible to have a completely new perspective on the use of Livonian vocabulary and Livonian dialects.

Taking all this into account, it becomes clear that the opportunities and technologies offered by the modern electronic world, when used wisely, can be a positive support for the preservation and development of all low-resource languages; and they are already helping to close the gap in resources between large and small languages.

6. Acknowledgements

This study was supported by the Latvian Ministry of Education and Science research program "Latvian language" sub-project "Livonian Language" (VPP-IZM-2018/2-0002).

7. References

Blumberga, R. (2006). *Lībieši dokumentos un vēstulēs.* Rīga: Latvijas vēstures institūta apgāds.

Bušs, O. (2015). Sõnaraamat või/ja eksperiment. *Keel ja kirjandus*, 10, pp. 744–746.

cl.ut.ee. *Tasakaalus korpus.* Accessed at: https://www.cl.ut.ee/korpused/grammatikakorpus/ (10 June 2019)

csb.gov.lv. *Centrālās statistikas pārvaldes datubāzes.* Accessed at: https://data1.csb.gov.lv/pxweb/lv/iedz/iedz___tautassk___taut___tsk2011/TSG11-06.px/table/tableViewLayout1/ (10 June 2019)

EKPS: *Eesti keele põhisõnavara sõnastik.* (2014). Tallinn: Eesti Keele Sihtasutus.

EKSS: *Eesti kirjakeele sagedussõnastik.* (2002). Tartu: Tartu Ülikooli Kirjastus.

ELD: *Eesti-läti sõnaraamat. Igauņu-latviešu vārdnīca.* (2015). Tallinn: Eesti Keele Sihtasutus.

Ernštreits, V. (2011). *Lībiešu rakstu valoda.* Rīga: Latviešu Valodas aģentūra, Līvõ kultūr sidām.

Ernštreits, V. (2012). Lībiešu valodas situācijas attīstība Latvijā. In I. Druviete (ed.) *Valodas situācija Latvijā: 2004-2010.* Rīga: Latviešu Valodas aģentūra, pp. 142−166.

Ernštreits, V. (2019). Electronical resources for Livonian. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages. Tartu: ACL SIGUR, the special interest group for Uralic Languages.* Accessed at: https://www.aclweb.org/anthology/W19-03 (30 July 2019)

LED: *Läti-eesti sõnaraamat. Latviešu-igauņu vārdnīca.* (2015). Tallinn: Eesti Keele

Sihtasutus.

LELD: *Līvõkīel-ēstikīel-leţkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca.* (2012). Tartu, Rīga: Tartu Ülikool, Latviešu valodas aģentūra.

LLLS: *Līvõkīel-leţkīel-līvõkīel sõnārõntõz. Lībiešu-latviešu-lībiešu vārdnīca.* (1999). Rīga: Līvõ kultūr sidām.

LW: *Livisches Wörterbuch mit grammatischer Einleitung.* (1938). Helsinki: Suomalais-Ugrilainen Seura.

lingua.livones.net. Accessed at: http://lingua.livones.net/lv/module/login (14 June 2019)

livones.net. *Lībiešu valodas vārdnīca.* Accessed at: http://www.livones.net/lili/lv/vardnica/ (10 June 2019)

murre.ut.ee. *Līvõkīel-ēstikīel-leţkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca.* Accessed at: http://www.murre.ut.ee/liivi/ (10 June 2019)

SW: *Joh. Andreas Sjögren's Livisch–deutsches und deutsch–livisches Wörterbuch.* (1861). St. Petersburg: Kaiserlischen Akademie der Wissenschaften.

unesco.org. *UNESCO Atlas of the World's Languages in Danger.* Accessed at: http://www.unesco.org/languages-atlas/ (13 June 2019)