Taking the Livonians into the Digital Space

Valts Ernštreits[1] and Gunta Kļava[2]

[1] University of Latvia Livonian Institute, Kronvalda 4-220, Riga, Latvia
valts.ernstreits@lu.lv
[2] University of Latvia Livonian Institute, Kronvalda 4-220, Riga, Latvia
gunta.klava@lu.lv

Abstract. The Livonians are a Finno-Ugric nation indigenous to Latvia. They are presently the most endangered culture in the European Union and their language is one of the most endangered languages in the world. The Livonian speech community lost its last compactly inhabited territory during the Soviet occupation of Latvia and since then has experienced a steady but continuous loss of Livonian language speakers. Today, only approximately 20 people are able to hold a conversation in Livonian. While presently living scattered across Latvia, with their language no longer used as a medium of daily interaction, the Livonian community has taken this challenging reality and turned it to its advantage in its approach to creating language resources from a limited set of human and archival resources. This article is devoted specifically to the subject of digital humanities in the service of endangered languages and cultures. Understanding the specifics of endangered and especially critically endangered languages and cultures, which actually form the majority of the world's cultural and linguistic diversity, is critically important both in building tailor-made digital resources for such communities and applying ready-made instruments for their benefit. This article offers a broader picture of the challenges and opportunities that developers of digital resources for critically endangered languages and cultures have to take into account, based on the Livonian experience. The article summarises and presents some of the findings and ideas deriving from the creation of digital resources and tools for the Livonian language and culture, approaches for overcoming a lack of resources and limited data issues, as well benefits and experiences that low resource – especially extremely low resource – cultures and languages can derive from the Livonian experience.

Keywords: Endangered Languages, Low Resource Languages, Digital resources, Livonian, Data Integration, Data Processing.

1 Introduction

The Livonians are a Finno-Ugric nation indigenous to Latvia. Currently, approximately 20 individuals are able to communicate in Livonian [1] and approximately 250 individuals gave their ethnicity as "Livonian" on the last Latvian national census [2]; however, the actual number of Livonians is considerably greater. Livonian heritage has had a significant, though understudied, role in the formation of modern Latvia, the Latvian

language, and the Latvian nation, and has also been important across the broader Northern European region.

The Livonian community has been able to preserve its identity and also its language up to the present day despite its small size (19th century – 2400 [3], mid-20th century – ca. 900 [4; 5] and the complicated history of the Livonian speech area. This history included the loss of the last compact Livonian-inhabited territory following the establishment of a border zone along the Baltic Sea in the area encompassing the Livonian villages at the start of the Soviet occupation after World War II [6]. As a result, the Livonians were scattered across Latvia and the world, and this fact continues to pose an added challenge for them. The same is true for the contents of various archives relating to Livonian heritage (language, folklore, folk cultural objects, etc.), which for historical reasons were mainly collected by researchers from abroad and are therefore currently stored at various institutions located in different countries [7].

Though the Livonian language is currently considered to be one of the world's most endangered languages and is listed in the UNESCO Atlas of the World's Languages in Danger as critically endangered [8][1], the study of Livonian and related topics requires the same opportunities and tools as those for any other language and culture. As a result, taking all of the aforementioned facts into account, the main issue faced by Livonian researchers, the Livonian community, and society at large, is ease of access to sources and archives relating to Livonian heritage. The rapid digitalisation of society during recent decades has created new opportunities for many smaller communities, including Livonians, to solve these problems and design new tools in order to ensure that these materials can be easily accessed and used. As with many other endangered languages and cultures, Livonian is under resourced. This poses additional challenges in any effort to preserve, sustain, or revitalise the Livonian language and culture, including the creation of digital resources and tools for this purpose.

This article is devoted specifically to the subject of digital humanities in the service of endangered languages and cultures. Understanding the specifics of such languages and cultures, which actually form the majority of the world's cultural and linguistic diversity, is critically important both in building tailor-made digital resources for such communities and applying ready-made instruments for their benefit. Our article is intended to offer a broader picture of the challenges and opportunities that developers of resources for critically endangered languages and cultures have to take into account, based on the Livonian experience. For this purpose we wish to summarise and present some of the findings and ideas, approaches for overcoming a lack of resources and limited data issues, as well benefits and experiences that low resource – especially extremely low resource – cultures and languages can derive from the Livonian experience.

2 Building Resources for Endangered Cultures

When considering digital technologies as an opportunity for all languages to survive (especially endangered ones), it is important to remember that for most languages in

[1] Although it incorrectly states that the last first-language speaker died in 2009.

the world, inclusion in the digital domain remains a distant possibility. However, the good news is that the same technologies that empower more widespread languages can also empower the less widespread – or even endangered – ones. Although for endangered languages, this work should be carried out with special effort and advocacy [9].

2.1 General Challenges

The general challenges to be taken into account when building digital resources for low-resource languages and cultures are widely known – insufficient institutional, financial, political support for research and development, a lack of human resources, and especially a lack of skilled personnel. Another challenge is limited data leading to fewer automation options when building digital resources and tools. However, when it comes to languages and cultures, which are extremely small in size, there are additional issues to be taken into account.

First of all, any resources compiled or collected in an endangered language with very few speakers cannot be effectively processed or used without overcoming the language barrier. While the involvement of skilled personnel is already limited for low-resource cultures, finding such personnel with sufficient proficiency in an endangered language limits any effort even further. One of the ways of overcoming this issue – as is done in the Livonian case – is to involve the speaker community, utilising them not only as language consultants, but also involving them in the creation, processing and revision of content. Such an approach of course helps to create valuable linguistic resources; however, supervision of this process by linguists and specialists to ensure quality of the material is a necessity [10].

This involvement means that resources have to be useful and beneficial not only for the researchers, but also for the language and culture community itself, which is an important target group of potential users for any such resources. It is important, however, that while in most cases resources for endangered cultures are created with an intent to help to ensure the sustainability, preservation, and development or even revitalisation of such cultures, such work is ineffective if carried out without close cooperation with the community and responding to its crucial needs.

For the Livonian language, this is evidenced by a long history of compiling Livonian dictionaries. The first major Livonian dictionaries were already published in the mid-19th century [11, 12], but due to the fact that they used phonetic transcription and German as a second language, the first dictionary actually usable by the Latvian-proficient Livonian community was only published in 1999 [13], coming almost a century too late for revitalisation efforts.

2.2 Challenges for Small Linguistic Communities

Trying to match the outputs of the resources to community needs can also present additional challenges. An example of this is a UL Livonian Institute project dedicated to language acquisition. During the course of this project, a survey was conducted to understand the Livonian-language acquisition needs in the community with an eye towards including these needs in tools based on resources developed for the Livonian language and culture.

Nicholas Ostler [14] identified four major focus areas regarding the role of technology in making the use of every language possible during this century. One of these areas – mechanisms to improve language access for members of a language community – is crucial for Livonian in its present situation, where even those speakers who want to maintain and simply use their language cannot do so, because of insufficient domains of language use. Consequently, the importance of digital solutions becomes greater and more urgent.

In studying the language acquisition needs (as well as other aspects of language acquisition) of the community, some general issues emerge relating to the theoretical facets of language acquisition, pedagogy, language standardisation as well as practical issues affecting the processes of language maintenance, development, and possibilities for use. In-depth interviews with good Livonian speakers (B and C language proficiency level) and questionnaires completed by other Livonian community members showed a particular interest and emphasis on appropriate digital solutions for language use or 'language on the computer' as one area where there was particular interest and hope for expanding the use of Livonian in real life and fostering language transfer to the younger generation. It is important to note that respondents used the concept 'language on the computer' to mean different digital language resources (from dictionaries and the possibility to check for a correct word form, to the expanded availability of language corpora, texts, information about and in Livonian, etc.) and less to language in other media, which is important for younger people, but must be further researched from the perspective of endangered languages and the effect on language practice [15]. The pandemic situation during the last year has forced all people around the world to decrease direct communication and contact; however, as a result, the idea of needing to make language easier to use for different purposes in the digital space has begun to make even more sense.

The complex nature of this process can be seen more clearly when analysing just one aspect of Livonian language revitalisation, namely creating language learning and teaching materials (for Livonian as a second language or L2, or as a heritage language). This task seems basic and straightforward from the theoretical perspective of larger languages used in daily life, but it becomes more complex when such materials have to be created for a language like Livonian, where challenges are not only related to linguistic characteristics but maybe even more to sociolinguistic issues.

Creating such materials reveals a variety of challenges: an absence of already existing pedagogical materials or an accompanying tradition of such materials to build on, a lack of pedagogically-trained personnel as well as insufficient language proficiency among potential teachers in general, a lack of language acquisition programmes for teachers and the wider community, an absence of certain domains of language use accompanied by an instant need for new terminology to fill these gaps, a lack of standards, which are central to creating learning materials, improving accessibility, and more.

Thus, the smaller a language or culture community, the more aspects actually have to be taken into account when developing even the simplest resources.

3 Digital Resources for Livonian

Although the first attempts to create digital resources for Livonian already took place in the 1990s (e.g. Livonian-Latvian-Livonian dictionary [13]), more active work began with the publication of the Livonian-Estonian-Latvian Dictionary [16] in 2012 and subsequent attempts to use data from this dictionary for creating various linguistic datasets and databases. This work led to the cluster of databases currently being further developed by the UL Livonian Institute.

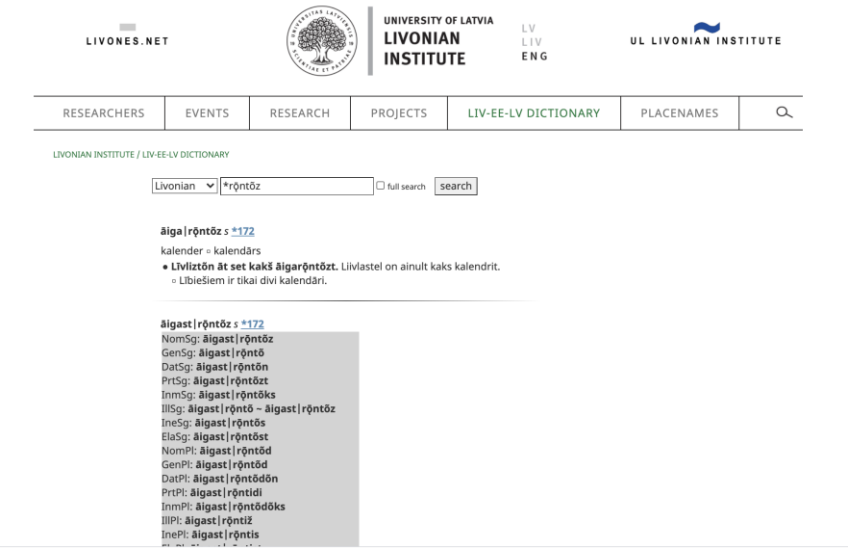

Fig. 1. Screenshot from the publicly accessible online dictionary.

3.1 Existing Resources

The cluster consists of three existing interconnected linguistic databases (lexical and morphological databases and a text corpus) and three databases in development (a data source database, an informant database, and a geospatially linked place name database). Completion of beta-testing of the entire cluster is planned for February 2021. The database cluster is built using a system developed in-house and is primarily targeted towards providing better data extraction from limited data sources as well as consistency of extracted data over the entire cluster. The database cluster is currently fully available with restricted access for database developers and researchers [17], and partially – for publicly open access in the form of a Livonian-Estonian-Latvian online dictionary with morphological data included [18] and Livonian place name registry [19].

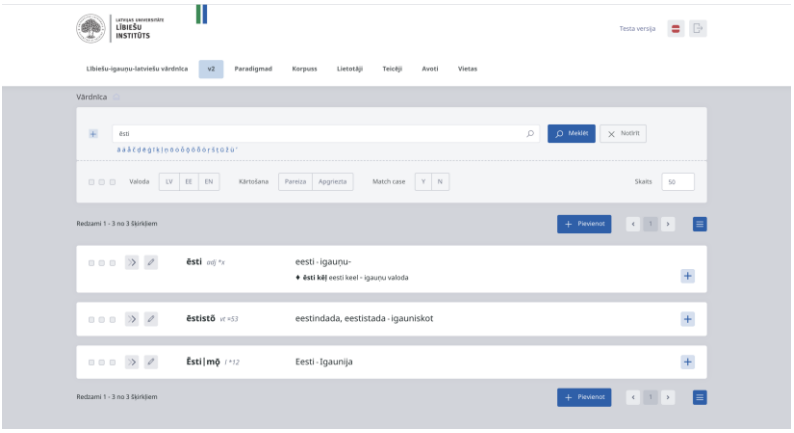

Fig. 2. Screenshot from the restricted-access lexicography module.

Creation of existing resources was initially entirely data and community driven. It started with the need to have an online Livonian dictionary which could be easily updated and expanded, serving not only as a collection of vocabulary, but also as a reference for the Livonian literary standard (currently the open-access dictionary of standardised Livonian contains ca. 12500 lemmas). Due to the fact that Livonian has extremely rich morphology, dictionary lemmas were then supplemented with a database of morphological forms (all forms and their variations of declinable words). The need to acquire additional vocabulary from texts, including textual parts of the Livonian-Estonian-Latvian Dictionary [16] itself (examples used to illustrate lemmas), led to the creation of the corpus (ca. 500 000 indexed and unindexed words), which utilises the indexation of lemmas and their morphological forms from the lexical and morphological databases, respectively.

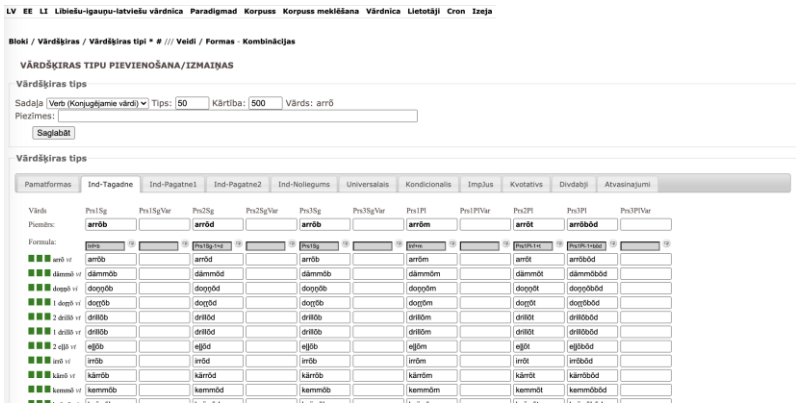

Fig. 3. Screenshot from the morphology module.

Since in many cases, endangered languages also have underdeveloped or no language standards at all, an important task for the Livonian linguistic databases is setting that standard; therefore, they also include standardising options. It is possible to indicate lemmas, morphological forms, and also texts (including parts of texts or even sentences) that may be considered as standardised and thus made available to the general public. Since the research, standardisation, collection, and processing of data happens simultaneously in most cases, these standards must also be dynamically adjustable.

Though still under development, these resources have already significantly simplified the process for learning about and studying the Livonian language and culture, and have created a foundation for future solutions directed towards simplifying access to Livonian-related materials. Existing resources have been serving as a base for contemporary research and subsequently also the standardisation of Livonian, and currently an overview of Livonian grammar is being prepared based on data acquired from these databases.

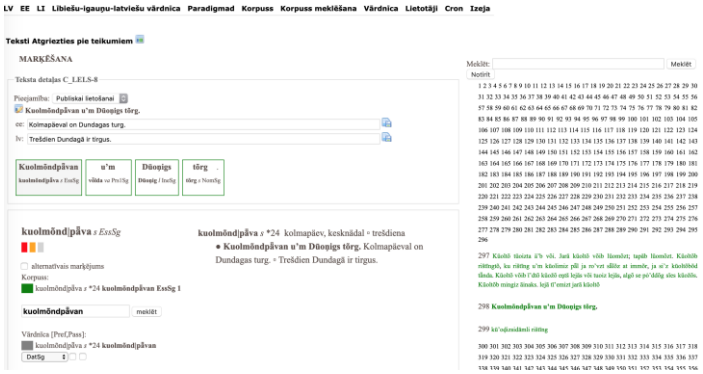

Fig. 4. Screenshot from the corpus module.

3.2 Resources Developed

In 2020, work started on further development of the cluster by adding three important datasets and expanding it with cultural data. As the corpus is not only a source for linguistic data, but also a collection of texts (folklore, culture texts, newspaper texts), the cluster is supplemented with background data for the sources it uses, with an option for the inclusion of data describing future sources such as ethnographic collections, etc. Similarly, background data on people (including language informants and owners of original items) is added. Both of these are also necessary from the perspective of achieving better quality and results from the research.

Limited data does not permit the use of statistics as a deciding factor in many cases, when research is performed in specific areas. In languages and cultures with sufficient data, statistics normally point towards the most common and regular use, but for endangered cultures, which have limited data and are regularly exposed to majority cultures, this may not always be the case. If, for example, two forms of a word are different out of a total of ten recorded, it might well be that these two are actually correct and the rest of the cases are mistakes caused by the specific backgrounds of each informant or

for other reasons. Therefore, data on the circumstances and authors of each source becomes more important, the smaller the community being researched. This, however, raises the question of how to ensure data anonymity when researching endangered cultures.

The third database currently under development – a geospatially linked place name database – is formally an extension of the lexical database. The creation of this database currently includes ca. 191 Livonian toponyms (over 1000 expected by the summer of 2021) and was initially also community driven due to practical considerations relating to collecting Livonian place names, so that bilingual road signs could be introduced in the territories historically inhabited by the Livonians.

Creating this resource has been quite challenging. There is no single Livonian place name archive and place names are primarily found in lexicographic publications or studies, albeit in a more scattered manner and fewer in number [12, 16, 20, etc.]. Currently, recording a large number of place names and localising them through field work is also not possible, as there remain only very few Livonian speakers and – due to the historical situation of the Livonians – their connection with the territories inhabited by their ancestors is often indirect and their knowledge of Livonian place names is meagre.

Thus, in large part, the Livonian language corpus is also used to acquire data for this resource as well as metadata from various collections that points to the place of collection, including folklore collections, ethnographic collections, etc. located in different institutions in Latvia and abroad. Mapping the place names is very important, as these places not only form a layer of linguistic data, but also reflect locations of specific importance to Livonian culture. These places are often not found in any available cartographic product (for example, homesteads which have disappeared or have been moved, drained rivers or lakes, etc.). Mapping itself is performed using several approaches, including textual study, comparison with corresponding Latvian place names that are in many cases borrowed or translated from Livonian, etc.

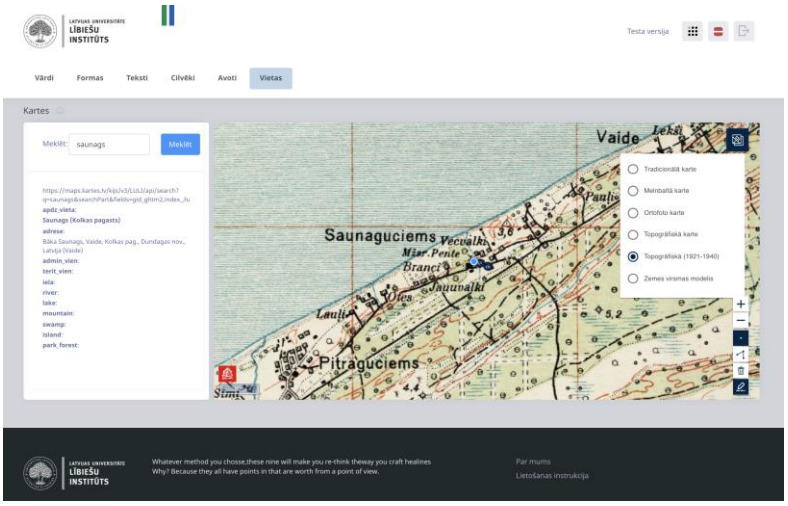

Fig. 5. Screenshot from the place name module (tool for adding the geospatial information).

This resource is particularly important in expanding possibilities for research and access to other Livonian language and culture resources. The mapped place name database is planned to serve as a starting point for the creation of a one-stop database cluster and its backbone. This will make it possible to link place names, geospatial data, and information from other fields (e.g., information on informants or original possessors of sources, dialect materials, objects, folklore, oral history collections, etc.), thereby opening up new possibilities for the multifaceted documentation and research of Livonian heritage in the future. Furthermore, this will make it possible to use data for many different purposes beyond research. These could include, for example, digital exhibits or using the cartographic products resulting from this project for education exploring regional heritage, cultural tourism, development of municipal strategies, entrepreneurship, and many other areas. The need for such outputs can already be seen in various numerous requests by state and municipal institutions, private entrepreneurs and NGOs that the UL Livonian Institute is trying to answer.

3.3 Data Consistency and Management of Human Resources

One of the benefits that endangered languages and cultures can have from possessing only limited data is data integration. While larger communities tend to develop their resources across several institutions, applying different principles and thus struggling with the need to link data among them, endangered cultures are capable of handling data in a more consistent way. This is a key data principle integrated into Livonian resources that also ensures better data extraction and handling later on, especially in light of insufficient human resources.

In resources developed for Livonian, each kind of data is kept separate, but is used in other datasets as linked data. Thus, the lexical database only contains data on lemmas and meanings, but no textual units – these are located in the corpus. Even example material found in the Livonian-Estonian-Latvian Dictionary – forms separate text in the corpus; examples in the current dictionary are featured as corpus links. The same applies to the place name database, which contains geospatial data, but the place names it features as lemmas are linked from the lexical database along with examples from the corpus and grammatical information from the morphological database. This approach ensures that adjustments or alterations made in one dataset are reflected correctly in other resources without any additional effort.

Another important – and inevitable – principle of Livonian-language resources is multilingualism. In order for Livonian intangible heritage (linguistic data, folklore, cultural texts, etc.) to be available to researchers, the Livonian community, and also the general public, which in large part has no proficiency in Livonian, translations have to be provided for almost every textual unit. The primary languages selected for translations are Latvian – as the main language used by the Livonian community and the general public in Latvia – and English for international users. However, as the Livonian language and culture has been documented and researched in a number of countries, there is also a point in using already existing translations in, e.g., Estonian, Finnish, German, etc.

The Livonian experience also shows that when developing digital resources for such tiny communities, manual work is inevitable and only some processes can be fully entrusted to automated solutions, at least in their initial phases. For example, 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 part of them are available electronically, automated indexing does not work, because of a lack of clear and verified grammar rules and limited data, etc. [21].

As there are also few sufficiently proficient to work with intangible heritage data in an endangered language, the question of how to acquire resources and increase the efficiency of existing ones becomes crucial. In the Livonian case, this issue is addressed by both technical and political solutions. On the technical side, this is done first by lessening the amount of work through the use of data handling and data extraction principles, e.g., when corpus data is indexed, it is also added to the morphological and lexical databases.

Resource efficiency is also solved by workflow matching different skill levels – those less skilled can perform simple tasks that tend to form the majority of data processing, while complex tasks are left for those more skilled. This comes along with a principle of providing access to data as soon as it is added to a database. This also ensures that dynamic adjustments can be made by those processing or using data for their research. Because of this approach, it is also possible to use crowd sourcing done by researchers, where those performing research using Livonian resources may contribute at every stage of their work by eliminating spotted mistakes, adding missing indexation, translation, and so on during the course of their research.

To help solve the problem of an insufficient number of personnel proficient in Livonian available to work on the resources, a political solution has also been suggested. The UL Livonian Institute is currently implementing a pilot project and has offered the idea to launch a programme for language speaker community involvement. In this programme, language speakers would be paid for working on various tasks involving language, including processing (indexing, transcribing texts) or adding data (e.g., voice data for the dictionary) to the digital resources under development.

4 Conclusions

The emergence of digital technologies provides novel possibilities for preserving and providing access to endangered language data. Up to now this data has been too inaccessible for linguists and remaining language speakers alike, and it is important that digital resources are created in close cooperation between both of these groups. As a result, there is mutual benefit by helping to ensure better sustainability for the endangered communities, while also significantly increasing the social impact of research results.

Although endangered cultures suffer greatly from a lack of resources, many issues can be resolved using strategic and carefully considered approaches to the structure of resources and workflows. In addition, sometimes even political solutions can be as important as technical features. At the same time, having limited data may be beneficial

for endangered cultures, as this may permit the creation of more integrated and consistent resources, thereby compensating for the lack of quantity with increased quality of data processing.

Though Livonian resources were initially created as a Livonian language data archive and a tool for language research, standardisation and acquisition, its principles have been expanded to create the cultural resources currently under development. In the future, these resources can also be further adjusted to suit other types of digital data such as images, audio recordings, videos, 3D scans, data from other databases, thereby offering a whole new perspective for studying and preserving one of the most endangered cultures in Europe.

The products and discoveries resulting from this resource may also be useful to other smaller communities, and the synergy and coordination among various resources can create a rich, high-quality resource suitable for multifaceted studies in many fields, or for interdisciplinary research in general.

Acknowledgements

This study was supported by the Latvian Council of Science Fundamental and Applied Research Programme project "Documenting and mapping Livonian place names and creating an official place name register" (LZP-2019/1-0240), the National Research Programme project "Digital Resources for Humanities: Integration and Development" (VPP-IZM-DH-2020/1-0001), and the PostDoc Latvia project "Applying the Common European Framework of Reference for Languages to Livonian: A New Opportunity for Endangered Languages" (1.1.1.2/VIAA/3/19/527).

References

1. Lībiešu valoda (Livonian language). Homepage, http://www.livones.net/lv/valoda/?libiesu-valoda, last accessed 2019/06/15.
2. Tautas skaitīšana 2011 (Census 2011). Centrālā statistikas pārvalde (2011). https://www.csb.gov.lv/lv/statistika/statistikas-temas/iedzivotaji/tautas-skaitisana/tabulas/tsg11-06/latvijas-pastavigie-iedzivotaji-pec, last accessed 2019/06/15.
3. Wiedemann, F. J.: Joh. Andreas Sjögren's Livische Grammatik nebst Sprachproben. Joh. Andreas Sjögren's Gesammelte Schriften. Band II. Theil I. St. Petersburg, Kaiserlischen Akademie der Wissenschaften (1861).
4. Mežs, I.: Jūrmalas ciemu iedzīvotāji 1935. gada tautas skaitīšanas uzskaites lapās. Lībiešu gadagrāmata, Līvlizt āigastrōntõz 2000, pp. 76–85. Lībiešu krasts, Valsts īpaši aizsargājamā kultūrvēsturiskā teritorija "Lībiešu krasts" (2000).
5. Blumberga, R.: Lībieši 19.–21. gadsimtā. In: Blumberga, R., Mäkeläinen, T., Pajusalu, K. (eds.). Lībieši. Vēsture, valoda un kultūra, pp. 169–204. Līvõ Kultūr sidām, Rīga (2013).
6. Druviete, I., Kļava, G.: The role of Livonian in Latvia from a sociolinguistic perspective. Eesti ja soome-ugri keeleteaduse ajakiri = Journal of Estonian and Finno-Ugric Linguistics 9(2), Livonian studies III, pp. 129–146. University of Tartu, Tartu (2018).
7. Ernštreits, V.: Lībiešu valodas situācijas attīstība Latvijā. In: Druviete, I. (ed.). Valodas situācija Latvijā: 2004–2010, pp. 142−166. Latviešu valodas aģentūra, Rīga (2012).
8. UNESCO Atlas of the World's Languages in Danger. Homepage, http://www.unesco.org/languages-atlas/, last accessed 2020/11/29.

9. Buszard-Welcher, L.: New Media for Endangered Languages. In: Rehg, K. L., Campbell, L. The Oxford Handbook for Endangered Languages, pp. 510–530. Oxford University Press (2018).
10. Drude, S., Ingason, A. K., Kristinsson, A. P., Arnbjörnsdóttir, B., Sigurðsson, E. F., Rögnvaldsson, E., Nowenstein, I. E., Sigurjónsdóttir, S.: Digital resources and language use: Expanding the EGIDS scale for language development into the digital domains. In: Conference "Communities in Control - Learning tools and strategies for multilingual endangered language communities". London; Minde (Portugal) (2018). https://www.researchgate.net/publication/335244251_Digital_resources_and_language_use_Expanding_the_EGIDS_scale_for_language_development_into_the_digital_domains, last accessed 2020/11/14.
11. Wiedemann, F. J.: Joh. Andreas Sjögren's Livisch–deutsches und deutsch–livisches Wörterbuch. St. Petersburg, Kaiserlischen Akademie der Wissenschaften, (1861).
12. Livisches Wörterbuch mit grammatischer Einleitung. Helsinki, Suomalais-Ugrilainen Seura. (1938).
13. LLLS = Līvõkīel-leļkīel-līvõkīel sõnārõntõz. Lībiešu-latviešu-lībiešu vārdnīca. Rīga, Līvõ kultūr sidām. (1999).
14. Ostler, N.: Endangered languages in the New Multilingual Order per genus et differentiam. In: Jones, M. C. (ed.). Endangered Languages and New Technologies, pp. 1–16. Cambridge University Press (2015).
15. Moriarty, M.: New roles for endangered languages. In: Austin, P. K., Sallabank, J. (eds.). The Cambridge Handbook of Endangered Languages, pp. 446–458. Cambridge University Press (2011).
16. Līvõkīel-ēstikīel-leļkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca. Tartu, Rīga, Tartu Ülikool, Latviešu valodas aģentūra (2012).
17. Livonian language and culture database cluster. Homepage, http://lingua.livones.net/lv/module2/vardnica/, last accessed 2020/20/12. (2020).
18. Livonian-Estonian-Livonian dictionary. Homepage, http://www.livones.net/lili/en/vardnica/, last accessed 2020/20/12. (2020).
19. Registry of Livonian place names. Homepage, http://www.livones.net/lili/en/vietvardi/, last accessed 2020/20/12. (2020).
20. Boiko, K.: Baltijas jūras somu ģeogrāfiskie apelatīvi un to relikti Latvijas vietvārdos (Balto-Finnic geographic words and their relicts in the toponymy of Latvia). Disertācija filoloģijas doktora grāda iegūšanai. Latvijas Universitāte (1993).
21. Ernštreits, V.: Electronical resources for Livonian. In: Pirinen, T. A., Kaalep, H.-J., Tyers, F. M. (eds.). Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages, pp. 184–191, Tartu (2019). Homepage, https://www.aclweb.org/anthology/volumes/W19-03/, last accessed 2020/10/09.