Baltic J. Modern Computing, Vol. 12 (2024), No. 4, 493-500
https://doi.org/10.22364/bjmc.2024.12.4.11

A Balanced Vocabulary Without a Balanced Corpus: The Livonian Case

Valts ERNŠTREITS

University of Latvia Livonian Institute

valts.ernstreits@lu.lv

ORCID 0000-0002-5323-8536

Abstract. This article, using the example of Latvia's indigenous Livonian language, explores approaches to enhance the consistency of endangered language documentation and lexicographical resources, with a particular emphasis on creating a balanced vocabulary. For critically endangered languages like Livonian, existing collections may be small and imbalanced, often reflecting outdated and limited sources. This presents significant challenges in accurately capturing the vocabulary essential for everyday communication and language acquisition. The article examines how frequency data from related language, such as Estonian in the case of Livonian, can be utilized to fill gaps and develop a more consistent and representative vocabulary.

Keywords: Livonian, Endangered languages, Lexicographical resources, Language documentation, Frequency data

1. Introduction

The sustainability of critically endangered languages depends on many factors, some of which include proper and systematic documentation and data collection, as well as the existence of sufficient digital resources, which set the foundation for the successful development and application of digital tools and technologies (Ostler, 2015).

One of the key outcomes of documentation is often the data, which can be elaborated into lexicographical collections. These collections serve as a base for creating dictionaries that are made available to those who speak and learn the language. Especially for under-resourced and endangered languages, dictionaries serve as essential tools for language acquisition and use, also providing a source of orthography and terminology, thus significantly contributing to the sustainability of the language.

This article explores approaches that, through the application of digital resources, may improve the consistency and content quality of both documentation and dictionaries. Based on experiences of creating a lexicography database and dictionary for Latvia's indigenous Livonian language (Ernštreits et al.,, 2024c), this article discusses what methods may be applied in the creation and consistency assurance of lexicographical publications and resources of critically endangered languages, which may either lack sufficient data or have existing data affected by various factors that prevent proper reflection of actual language use and needs.

While formed and spoken in Latvia, belonging to the same Finnic language group as Estonian, Livonian is a critically endangered language with around 20 fluent speakers, though there is a growing community of learners involved in language acquisition (Moseley, 2014; Ernštreits, 2019, 2023; Laakso, 2022). Although significant efforts for documenting Livonian have taken place since the mid-19th century, these efforts decreased toward the end of the 20th century. As interest in acquiring and using Livonian grows (Kļava, 2021), there is an increasing need for lexicographical collections covering both basic vocabulary and terminology necessary for contemporary everyday use (Ernštreits and Kļava, 2023).

2. Documentation and Creation of Dictionaries for Critically Endangered Languages

One of the major flaws in documenting and publishing the vocabulary of critically endangered (and other lesser-used) languages like Livonian is often data-driven content. It is not unusual for language documentation to occur unevenly across different times and spaces, often being scarced and based on the specific research needs of the person conducting the documentation (frequently a researcher from outside the community) (Mosel, 2004). As a result, the purpose of the documentation may not be to serve the needs of the community, e. g. language acquisition or use.

In the creation of dictionaries, it is also not unusual for them to rely primarily on lemmas found in particular data collections, rather than actively seeking the basic and specific vocabulary needed by learners and speakers of the language. This can lead to significant inconsistencies in lexicographical publications, where even the most common everyday words may be absent, simply because they are not present in the corpus of documented texts or have skipped the attention while documentation. Another issue for lexicographic publications based on pre-existing documentation is that the sources used may be outdated and not meet the needs of contemporary users, especially if the documentation was completed earlier.

In the Livonian case, the Livonian-Estonian-Latvian Dictionary (LELD, 2012), which contains approximately 12,000 lemmas, exemplifies this issue. While this dictionary is the most comprehensive resource of Livonian vocabulary, it is uneven in coverage. Some domains of the language are well-represented, while others are almost absent. For instance, it lacks many basic words essential for everyday communication, such as the names of months or pleasantries like li[1] *pōlaks* (‘please’) and li *tienū* (‘thank you’). Even the term li *līvli* (‘Livonian person’) was missing from the dictionary and was only added during the final stages of its preparation. While this dictionary was published in 2012, it relies largely on data documented in the 1960s and 1970s. Consequently, it lacks not only modern terminology, such as ‘selfie’ or ‘zoom’, but also everyday terms like ‘sink’, ‘pizza’, ‘sticker’, and many more.

In recent years this dictionary has been transformed into a lexicography database (Ernštreits et al., 2024b) and publicly available multimodal dictionary[2], which is being

[1] Hereafter, the following language abbreviations are used: *li* – Livonian, *ee* – Estonian.
[2] www.livonian.tech

instantly expanded and equipped with additional data – morphology, English translations, audio, and geospatial data. However, issues of consistency and suitability for contemporary use become crucial as the dictionary is more and more used for community daily needs and language acquisition.

This leads to a key research question: What approaches can ensure the consistency of vocabulary collections and systematic documentation to maintain such consistency? Moreover, how can this be achieved when existing collections are unbalanced and contain only limited data, as is the case for many under-documented languages like Livonian?

In the digital age, as languages and language data are increasingly moving online, a clear answer seems to lie in the use of digitally extractable frequency data. But what can be done when unbalanced documentation results in frequency data that offers a distorted picture of actual language use?

3. Experiences Using Frequency Data for Estonian–Latvian Dictionary

In 2012, an ambitious initiative to create an Estonian-Latvian dictionary (Ernštreits et al., 2015) containing at least 40,000 lemmas was launched. Within an extremely tight timeframe—just 2.5 years. Before the project began, two key questions arose: how to obtain a consistent set of headwords that reflect contemporary Estonian use, and how to ensure the consistency and balance of the vocabulary in case the dictionary's compilation process had to stop before reaching the planned 40,000 headwords.

To address this, the project group suggested using contemporary frequency data and compiling the dictionary from core vocabulary (the most common and complex headwords) to peripheral vocabulary (seldom-used headwords). For this list of lemmas along with frequency data from several, somewhat overlapping sources – the Estonian basic vocabulary dictionary (Kallas et al., 2014), the Estonian frequency dictionary (Kaalep and Muischnek, 2012), a balanced corpus of Estonian (EKK) was obtained, and data was categorized in frequency groups.

The approach ensured that even in the early compiling stages dictionary was already covering the essential vocabulary of Estonian and kept expanding evenly, based on the actual use. But can such an approach be used for languages like Livonian?

4. Livonian frequency data

Technically, Livonian also has digital sources that can be used to obtain frequency data. The Livonian language and culture resource platform *Livonian.tech* (Ernštreits et al., 2024a) developed by the UL Livonian Institute also includes a corpus of written Livonian texts (Ernštreits et al., 2022). Compared to surrounding languages, it is relatively small – around 450,000 words, and has a lot of historical orthographical and dialectal variation due to very diverse texts (e.g., CLF, (ŪT, 1942), (Rätsep, 1959), (Mälk, 1980) etc.) it contains.

This corpus, however, is also heavily imbalanced. Approximately 30% of the corpus consists of religious texts, including a translation of the New Testament, while another 30% comprises folk tales. The remaining texts include dictionary examples, textbooks,

and other educational materials. The majority of the texts included in the corpus were collected or created in the 1930s, meaning the vocabulary may not reflect contemporary language proficiency and may lack modern notions.

While frequency data from the corpus may reveal some parts of the most frequent basic vocabulary and contribute to the overall expansion of data, it would still offer a distorted view of both general and contemporary language use.

5. Using Frequency Data from Other Languages

If the direct use of frequency data is not feasible for ensuring consistency, there may still be options to supplement it with frequency data from other languages. However, several aspects need to be taken into account.

The first is the relevance of the data. For instance, legislative and administrative systems in English-speaking countries or Scandinavia differ significantly from those in the Baltics. As a result, terminology and frequency data from one region may not align with actual language use in another. This suggests that using data from culturally and contextually closer languages may be more effective.

Another issue is grammar. For example, prefix verbs, which are extensively used in Latvian (an Indo-European language), typically correspond to adverbial constructions in Estonian and Livonian (Uralic languages). From this perspective, grammatically closer languages are likely to provide more precise data.

The third issue is the language proficiency of the dictionary compilers themselves. Even if frequency data from another language is highly relevant, it is of little use if the compilers lack proficiency in the particular language pair (e.g., Livonian-Estonian or Livonian-Latvian). Finally, there is the crucial question of the availability of aligned digital data, which is essential for the effective application of frequency data.

6. Managing the consistency for Livonian using Estonian data

In 2023, work began on preparing the reverse Estonian-Livonian dictionary in collaboration with the Estonian Language Institute. As part of this initiative and a state research program project[3], a test of the cross-referencing frequency data was conducted by comparing Estonian data from the *Livonian.tech* lexicography database with a list of 5,000 lemmas from the Estonian basic vocabulary dictionary (Kallas et al., 2014). For lemma identification purposes, this list was also supplemented with explanations of meanings from the Estonian Explanatory Dictionary (Langemets et al 2009).

The comparison involved exporting the list of Estonian correspondences, split by individual headwords (e.g., li *āigavait*: ee *ajavahemik*, *intervall* ‘interval’ > 1) ee *ajavahemik*; 2) ee *intervall*), and then aligning it with the headwords from the Estonian basic vocabulary dictionary. Three types of correspondences were sought after: *null* (no mention of a headword, e.g., ee *detsember* ‘December’); *exact* (headword mentioned as the only or first correspondence, e.g., ee *elus* ‘alive’: li *jelsõ* ‘alive’); and *like* (headword

[3] “Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia”, Nr. VPP-IZM-DH-2022/1-0002

mentioned as a secondary correspondence; e.g., ee *algaja* 'beginner': li *irgiji* ee *alustaja, algaja*). The results showed that out of 4,902 Estonian basic lemmas, 2,497 had exact correspondences, 568 were partially matching, and 1,837 were not present at all.

The next step was identifying correspondences for Estonian headwords not found in the Livonian data and reviewing partial matches. Through manual screening of the missing headwords, several main reasons for their absence in the data were identified: 1) headword missing, but the concept exists as a synonym (e.g., ee *kogus* 'amount': li *pāgiņõm* (ee *hulk*)), or the headword exists, but the particular meaning has not been included in the dictionary; 2) headword missing as it is expressed differently in Livonian (e.g., ee *ehkki* 'although': li *laz kil*); 3) headword is a derivative of an existing headword and not included in the dictionary (e.g., ee *ostmine* 'purchasing': li *võstõ* 'to purchase'; ee *kurvalt* 'sadly': li *murāgli* 'sad'); 4) headword is a compound in Estonian, but not in Livonian (e.g., ee *perekonnaliige* 'family member': li *aim nõtkõm*); 5) headword is only relevant for Estonian (e.g., ee *haigekassa* 'institution for medical insurance'; ee *võidupüha* 'Estonian Victory Day') or is semantically bound to Estonian (e.g., ee *pill* 'musical instrument'; ee *käesolev* 'this particular one': li *se* 'this'); 6) headword (or concept) has not been registered in the dictionary and needs to be located or created (e.g., ee *ketšup* 'ketchup'; ee *komm* 'candy').

Last group comprised from ca 1500 headwords (almost 1/3 of the headwords compared) and included both – core vocabulary (ee *lomp* 'puddle'; ee *puit* 'timber'; ee *pidulik*; ee *erinema* 'to differ'; ee *edukas* 'successful') as well as contemporary vocabulary (ee *teler* 'TV'; ee *eraisik* 'private entity'; ee *parool* 'password'; ee *ridaelamu* 'row house'; ee *limonaad* 'lemonade').

The Livonian dictionary was updated accordingly. Missing synonyms or meanings were added to headword entries, new headwords and terms were either located in the data or created, and Estonian headwords irrelevant to Livonian were excluded. Additionally, the list of lemma types in the database was expanded to include terms consisting of two or more headwords and constructions with specific meanings in Estonian (and also Latvian), including compound verbs and expression verbs.

7. Conclusions

Now that the work of updating the dictionary from Estonian data is nearing completion, it can be concluded that the approach of using the Estonian core vocabulary to identify missing concepts, gaps and flaws in the Livonian dictionary has proven to be highly beneficial. This approach has not only facilitated focused efforts on updating and improving the dictionary itself but has also provided valuable opportunities for data-driven language planning and vocabulary development. It has also offered language developers a list of crucial headwords and concepts to be standardized or created for contemporary Livonian daily use and language acquisition.

Estonian, which belongs to the same Finnic language family as Livonian, has been particularly useful in this regard. The two languages share many similarities in grammar, syntax, vocabulary and derivation, partially also regional cultural background, making Estonian a valuable reference point. It can, however, be observed that differences in grammar and semantic handling (e.g., compounding, derivational differences, and the use of different grammatical structures to express concepts) still play an important role, even between these two relatively close languages, complicating the identification of

missing concepts. This raises the question of whether using similar frequency lists from more distant languages would be equally efficient.

To test this, the next step the project team intends to take is to conduct a similar test using data (e.g., frequency data from the Korpuss.lv platform) from Latvian, the dominant contact language for Livonian and the language of the contemporary Livonian community. Latvian is important not only as a language that has significantly influenced Livonian but also as its closest cultural counterpart and a primary mediating source of modern international vocabulary.

8. End notes

When documenting a language with an uneven corpus and limited resources, it is crucial to employ methods that ensure the consistency and comprehensiveness of the vocabulary. Having basic contemporary vocabulary included is essential for several reasons. It provides a foundation for language learners, ensuring they acquire the most commonly used and culturally significant words necessary for everyday communication. It also preserves cultural heritage by ensuring the vocabulary reflects the traditional knowledge, practices, and values of the language community. Additionally, a balanced vocabulary supports linguistic research and the development of digital resources, such as dictionaries and language learning applications, which are crucial for the revitalization efforts of endangered languages.

Understanding and obtaing basic vocabulary for languages like Livonian with irregular documentation and unbalanced corpora is a challenging yet vital task. By leveraging frequency data from related languages (like Estonian in Livonian case) it is possible to ensure that the vocabulary is both comprehensive and relevant for modern use. This, in turn, supports broader language preservation and revitalization efforts, helping to ensure that Livonian remains a vibrant and living language for future generations.

For the language that needs to be documented or needs a lexicographic publication to serve the community needs of language use and acquisition, having a list of core vocabulary in some other language, preferably one being related in some way, is beneficial, even if direct cross-referencing or alignment is not possible. Such a list may help to guide the documentation process, ensuring that it is more systematic and focused on the most relevant and frequently used aspects of the language. Thus, approaches outlined in this article, which combine data to create a balanced vocabulary, offer perspectives for more systematic documentation and revitalization of endangered and contested languages, especially those that have little, imbalanced, or outdated documentation.

Acknowledgements

This article has been prepared as part of the State Research Programme project Nr. VPP-IZM-DH-2022/1-0002 “Towards Development of Open and FAIR Digital Humanities Ecosystem in Latvia”, implemented within the framework of the National Research Programme “Digital Resources of the Humanities.”

References

CLF = Loorits, O. *Collection of Livonian folktales*, Museum of Literature of Estonia, LF IV 6; ERA-13302-70814-18895

EKK = *Eesti keele koondkorpus* (*Estonian Reference Corpus*; in Estonian), Tartu, Tartu Ülikooli arvutilingvistika uurimisrühm. http://www.cl.ut.ee/korpused/segakorpus.

Ernštreits, V. (2019). Lībiešu kultūrtelpa / The Livonian cultural space, *Nemateriālais kultūras mantojums Latvijā – Nacionālais saraksts / Intangible Cultural Heritage in Latvia – National Inventory*, Rīga, Latvijas Nacionālais kultūras centrs, pp. 102–109.

Ernštreits, V. (2023). Lībiešu valoda (*The Livonian language*; in Latvian), *in* Ščerbinskis, V. (ed.), *Nacionālā enciklopēdija* (in Latvian). https://enciklopedija.lv/skirklis/5259-l%C4%ABbie%C5%A1u-valoda

Ernštreits, V., Muzikante, M., Grīnberga, M. (2015). *Igauņu-latviešu vārdnīca = Eesti-läti sõnaraamat* (*Estonian-Latvian dictionary*; in Estonian and Latvian), Latviešu valodas aģentūra, Eesti Keele Sihtasutus.

Ernštreits, V., Fišel, M., Rikters, M., Tomingas, M., Tuisk, T. (2022). Language resources and tools for Livonian. *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 13(1), pp. 13–36. https://doi.org/10.12697/jeful.2022.13.1.01

Ernštreits, V., Kļava, G. (2023). Chapter 4, Experiences in Teaching an Endangered Language: Finding the Motivation and Means to Ensure the Acquisition of Livonian, *in* Riitta-Liisa Valijärvi and Lily Kahn (eds.), *Teaching and Learning Resources for Endangered Languages*, Leiden, The Netherlands, Brill, pp. 66–82. https://doi.org/10.1163/9789004544185_006

Ernštreits, V. (ed. in chief), Vāvere, S., Viitso, T. R., Damberg, P., Kurpniece, M., Kļava, G., Balodis, U., Tuisk, T., Kūla, G., Tomingas, M., Soosaar S. E., Sedláčková, A., Jurgenovskis, T. (2024a). *Livonian language and culture resource platform "Livonian.tech"*, Riga, University of Latvia Livonian Institute. https://livonian.tech/

Ernštreits, V. (ed. in chief), Viitso, T.-R., Vāvere, S., Damberg, P., Kurpniece, M., Balodis, U., Tuisk, T., Kūla, G., Tomingas, M., Soosaar, S.-E., Sedláčková, A., Jurgenovskis, T. (2024b). *Livonian lexicographic database*. Riga: University of Latvia Livonian Institute. https://livonian.tech/

Ernštreits, V. (ed. in chief), Viitso, T. R., Kurpniece, M., Vāvere, S. (2024c). *Livonian morphology database*, Riga, University of Latvia Livonian Institute. https://livonian.tech/

Kaalep, H. J., Muischnek, K. (2012). *Eesti kirjakeele sagedussonastik* (*Frequency dictionary of Estonian*; in Estonian), Tartu, TÜ kirjastus.

Kallas, J., Tiits, M., Tuulik, M. (2014). *Eesti keele põhisõnavara sõnastik* (*Dictionary of Estonian Basic Vocabulary;* in Estonian), Tallinn, Eesti Keele Sihtasutus. https://arhiiv.eki.ee/dict/psv/

Kļava, G. (2021). The Effect of Covid-19 on Livonian Language Learning Opportunities. *Multiethnica*, *Journal of the Hugo Valentin Centre*, *41*, pp. 88–99. https://doi.org/10.33063/diva-472017

Laakso, J. (2022). *The Oxford Guide to the Uralic Languages*, ed. by Bakró-Nagy et al., Oxford University Press.

Langemets, M., Tiits, M., Valdre, T., Veskis, L., Viks, Ü., Voll, P. (2009). *Eesti keele seletav sõnaraamat* (*The Explanatory Dictionary of Estonian;* in Estonian),. Tallin, Eesti keele sihtasutus. https://arhiiv.eki.ee/dict/ekss/

LELD (2012). *Līvõkīel-ēstikīel-leţkīel sõnārõntõz. Liivi-eesti-läti sõnaraamat. Lībiešu-igauņu-latviešu vārdnīca*. (*Livonian-Estonian-Latvian Dictionary*; in Livonian, Estonian and Latvian). Tartu, Tartu Ülikool, Rīga, Latviešu valodas aģentūra.

Mälk, V. (ed.) (1980). *Liivi vanasõnad. Eesti, vadja ja läti vastetega (Livonian proverbs with Estonian, Votic and Latvian correspondencies; in Estonian)*, I–II, Tallinn Kirjastus "Eesti raamat".

Mosel, U. (2004). Dictionary making in endangered speech communities, *Language Documentation and Description 2*, pp. 39–54. https://doi.org/10.25894/ldd289

Moseley, C. (2014). Livonian – the most endangered language in Europe? *Eesti Ja Soome-Ugri Keeleteaduse Ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 5(1), 61–75. https://doi.org/10.12697/jeful.2014.5.1.04

Ostler, N. (2015). Introduction: Endangered languages in the New Multilingual Order *per genius et differentiam, in* Mari C. Jones (ed.), *Endangered Languages and New Technologies*, Cambridge, Cambridge University Press, pp. 1–13.

Rätsep, H. (1959). Liivi fraseoloogiat (*Livonian phraseology;* in Estonian). *Emakeele Seltsi Aastaraamat*, Tallinn, pp. 226–242.

ŪT (1942). *Ūž testament* (*New Testament*; in Livonian). Helsinki, Herättäjä-yhdistys.

Received November 25, 2024, accepted November 28, 2024