Towards standardized inflected lexicons for the Finnic languages

Jules Bouton
Université Paris-Cité, LLF, CNRS
jules.bouton@u-paris.fr

Abstract

We introduce three richly annotated lexicons of nouns for Livonian, standard Finnish and Livvi Karelian. Our datasets are distributed in the machine-readable Paralex standard, which consists of linked CSV tables described in a JSON metadata file. We built on the morphological dictionary of Livonian, the VepKar database and the Omorfi software to provide inflected forms. All noun forms were transcribed with grapheme-to-phoneme conversion rules and the paradigms annotated for both overabundance and defectivity. The resulting datasets are usable for quantitative studies of morphological systems and for qualitative investigations. They are linked to the original resources and can be easily updated.

1 Introduction

1.1 Rationale

Over recent years, the amount of morphological resources available for the Uralic languages has strongly increased. Reasons for this are (a) the efforts of Finno-Ugrists to provide dictionaries and translation tools for minority languages; (b) the interest of typologists for computational approaches to linguistic diversity. However, these resources are scattered across different standards and do not necessarily fit the needs of morphologists. Although recent researches in computational morphology rely on various approaches (e.g. Malouf, 2017; Baayen et al., 2019; Beniamine et al., 2021), they all share the need for high quality morphological data in phonemic transcription.

Several projects strive to provide good coverage of the numerous Finnic languages. Recently, lexicons following the UniMorph format have flourished: Finnish (Kirov et al., 2016); Estonian and Northern Sami (Kirov et al., 2018); Livvi, Livonian and several other (McCarthy et al., 2020); Võro (Batsuren et al., 2022). Despite its increasing size, Malouf et al. (2020) have shown the pitfalls of UniMorph when it comes to linguistically informed studies of morphological variation. Semantic information, inflectional classes or frequencies are hard to extract and wordforms are provided in orthographical representations. The GiellaLT infrastructure (Pirinen et al., 2023) also provides acccess to dozens of morphological rule-based parsers. However, they are intended to enhance language-learning tools and they are not meant for morphological investigation either.

On the other hand, scholars and language institutes have developed their own resources, providing both inflected forms and rich annotation. Such resources are invaluable, but there is few of them. As a result of their dispersal, they are provided in different formats and through idiosyncratic infrastructures which make them less accessible for large scale comparative studies. Still, efforts for interoperability exist: in UniMorph 3.0, resources for Karelian languages are directly extracted from the VepKar database (McCarthy et al., 2020), although a lot of information is lost in the conversion, due to the limits of the UniMorph format.

Our lexicons in phonemic transcription are designed to fill this gap. We selected valuable, well-curated and rich resources for three Uralic languages from the Finnic group with very different backgrounds. Standard Finnish is the national language of Finland, spoken by around five million people in Finland.[1] Livvi Karelian is a southern Karelian language spoken by 25,000 individuals in Russia, near lake Ladoga. Courland Livonian is a minority language spoken until the end of the 20th century on the coast of Courland. Although our pipeline can in theory be extended to verbs, this release only covers nouns. As our main contribution, we enriched the datasets with phonemic transcriptions and linguistic annotations.

[1] Statistics are from the corresponding chapters of Bakró-Nagy et al. (2022).

Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages, pages 59–66
November 28-29, 2024 ©2024 Association for Computational Linguistics

Dataset		ISO	Licence	DOI	Cells	Lexemes	Forms
ParaLiv	1.0	liv	CC BY-SA 4.0	10.5281/zenodo.11391421	16	6,769	110,449
ParaKar	1.0	olo	CC BY-SA 4.0	10.5281/zenodo.13736171	33	4,975	196,555
ParaFin	1.0	fin	GNU GPL v3	10.5281/zenodo.13736132	151	5,000	879,117

Table 1: Main properties of the three datasets

1.2 The Paralex format

Beniamine et al. (2023) introduced the Paralex standard[2], which provides a structured way of representing morphological data. A Paralex dataset is a relational database constituted of CSV files linked together by relations. Beniamine et al. (2024) provide a detailed presentation of the structure of such a dataset. Thanks to the underlying Frictionless framework (Fowler et al., 2017), a Paralex dataset is adaptable to one's needs but also machine-readable. Thus, the Paralex standard puts good data management practices (FAIR: Wilkinson et al. 2016 ; DEAR: Beniamine et al. 2023) at the core of the dataset development.

Paralex datasets are intended for morphologists. As such, they offer two crucial improvements over other formats: phonemic representations and rich annotations. Since orthographic representations of words often obfuscate crucial features, the inflected forms are provided both in orthographic and phonemic writing. The phonemic transcriptions are checked on development sets to avoid regressions and cover most of the morphologically meaningful contrasts. Allophony is left out when it doesn't affect morphology. Paralex takes into account morphological diversity and has built-in methods to tag variants or defectivity (see below).

Our datasets follow these principles. They are made available on Zenodo under the names *ParaKar*, *ParaFin* and *ParaLiv* (see Table 1). The pipelines used to build the lexicons are available on Gitlab and ensure replicability of the results. Changes in the upstream sources can easily lead to updated versions of the datasets thanks to Zenodo's versioning system. They are distributed under open-source licences.

2 Building the lexicons

2.1 Lexemes and forms

For Livonian, we relied on the morphological dictionary of the Livonian Institute (Ernštreits et al., 2024), which itself builds on the Livonian dictionary by Viitso and Ernštreits (2012). In the absence of reliable frequency information, we provide support for all the nouns in the dictionary. We extracted the inflected wordforms and their properties as a JSON file and controlled the quality of the forms. A dozen of lexemes required upstream corrections and were ruled out. All the cells available in the dictionary were retained, which does not include lexicalized external local case forms. In compounds, the boundary between the components is marked. For phonological reasons, the derivatives ending in *-nikā* were treated as compounds, following Posti (1942, 301).

Similarly to Paralex datasets, the VepKar corpus used for Karelian (VepKar, 2009/2024; Boyko et al., 2022) is a relational database with annotated tables. Thus, converting the extracted tables was rather straightforward, despite the difference in the data structure (resp. CSV and MySQL). As for Livonian, the VepKar database provided pre-inflected forms for Livvi (Novak et al., 2020; Krizhanovskaya et al., 2024). Since VepKar has a better support for New Written Livvic, we focused on this variety of Livvi and excluded forms from other dialects. We retained all the lexemes that were attested at least once in the corpus. We replaced the accusative cell used in VepKar by genitive and nominative labels, depending on the form in question.[3] We additionally filtered the database and corrected a few forms. VepKar features an affix column, which made it possible to insert a boundary in wordforms after the immutable part of the stem. Compounds are segmented.

The situation of Finnish is different as we did not use a database of wordforms. We selected the 5000 most frequent nouns from the frequency dataset provided with the LASTU software (Itkonen

[2] https://paralex-standard.org/

[3] With respect to the accusative, the situation in Karelian is similar to that in Finnish. Bielecki (2009) shows that older descriptive grammars introduced an accusative while recent accounts only feature nominative and genitive. While syntacticians tend to agree in favour of an accusative (Holmberg and Nikanne, 1993), we adopt here a morphological perspective.

et al., 2024), which in turn relies on the Finnish Parsebank (Luotolahti et al., 2015). We then matched those nouns with the internal resources of the Omorfi HFST (Pirinen, 2015; Pirinen et al., 2017) and used the generator to produce inflected forms. Although the interaction of clitics, case and number markers and personal suffixes leads to a large amount of paradigm cells, we decided to only retain the combination of case, number and possessive suffixes. In our dataset, this already amounts to 151 cells.[4] Compounds and immutable stem boundaries are marked as well.

Table 1 summarizes the quantitative properties of the extracted datasets. All have around 5000 lexemes, which is a standard size for such resources (Beniamine et al., 2024).

2.2 Phonemic transcriptions

Grapheme-to-phoneme (G2P) transcription was performed with the Epitran software (Mortensen et al., 2018). Epitran requires a mapping of graphemes to phonemes and a set of pre- and post-processing regular expressions. For our datasets, we used a bundle of custom and modified rules.

For Livonian, we used a heavily modified version of the Estonian rules built for the Eesthetic package (Beniamine et al., 2024). Traditional accounts of Livonian phonology (Posti, 1942; Viitso, 2007) introduced numerous distinctions which are not always crucial for a phonemic description. For our transcription we relied on Tuisk's (2016) analysis and complemented it with previous accounts. We review the most crucial design choices.

Traditional accounts distinguish between short phonemes, long phonemes, short geminates and long geminates. We decided to keep a three-fold distinction for consonants and a two-fold opposition for vowels (ex 1). Due to the existence of feet isochrony (Viitso, 2007, 49), we mark the length of the first syllable coda when the second syllable is short (ex 2). Livonian is known for its tonal opposition (broken or plain) which affects accented syllables (Tuisk, 2015). We transcribe the broken tone as a property of vowels and polyphthongs and mark it with a superscript glottal stop ˀ (ex 3). We insert glides where required before orthographic <ž>, <j> and <v> (ex 3). Finally, Livonian displays a wide range of polyphthongs which were all documented. Table 2 showcases the triphthongs.

[4]For possessives, the values 3SG and 3PL are treated as syncretic. For instance, the cell NOM.SG.3 covers singular and plural possessors.

	Front-back		Back-front	
	Plain	Broken	Plain	Broken
All short	ieu	ieˀu	uoi	uoˀi
Last long	ieuː		uoiː	
First long			uːoi	uːˀoi

Table 2: Inventory of Livonian triphthongs found in our dataset

(1) a. kik → kikː 'rooster' NOM.SG
b. kikīd → kikˑiːd NOM.PL
c. kikkõ → kikːɯ PART.SG

(2) a. mustā → musˑtaː 'black' NOM.SG
b. mustõ → musːtɯ PART.SG

(3) ke'ž → ke̯ˀiʒ 'flea'

For Finnish, we used a modified version of the Finnish G2P converter introduced in Epitran 1.25. We don't mark the allophones of /h/, /s/, /l/, /m/, /n/ (ex 4), but we added additional rules to distinguish diphthongs from vowel sequences (ex 5) in conformity with Suomi et al. (2008, 49-51). We marked as a glottal stop the stop that alternates with intervocalic /k/ during gradation (ex 6). Following Karlsson's (1983, 349) view, morphs triggering boundary lengthening were not considered in the phonemic transcription, but we documented them in the analysed orthographic and phonemic transcriptions with the superscript symbol ˣ (ex 6).

(4) a. vihko → ʋihko 'notebook'
b. kohta → kohtɑ 'place'

(5) a. hyötyä → hy͜øtyæ 'benefit'
b. aie → ɑ͜ie (ɑ͜ieˣ) 'intention'

(6) vaa'an → ʋɑːˀɑn 'scale'

The Karelian G2P is a slightly modified version of the Finnish one. It is based on Pyöli (2011), but was extended with more detailed sources (Novak et al., 2022; Arhimaa, 2022). The Livvi transcription covers the digraphs and affricates specific to Karelian (ex 7) and introduces support for the contextual palatalization of /l/, /n/, /r/, /d/ and /t/ (ex 8) following the principles described by Novak et al. (2022, 58). We included palatalized and voiced geminates and we took into account the existence of six triphthongs, although they do not occur in our dataset as they are limited to verbs.

(7) čondžoi → t͡ʃond͡ʒo͜i 'flea'

(8) ellendys → elʲːendys 'wisdom'

–	Our datasets	`ill.sg`
–	UD	`N;IN+ALL;SG`
–	UniMorph	`Case=Ill|Number=Sing`
fin	Omorfi	`[NUM=SG][CASE=ILL]`
olo	VepKar ID	`10`
liv	Liv. Institute	`IllSg`
liv	Tartu	`sg.ill.`

Table 3: Mapping of the ILL.SG cell to other dialects

3 Rich annotations

Phonemes and graphemes For each dataset, we provided a grapheme inventory to ensure consistency in our orthographical sources. All three datasets also contain a machine-readable phoneme inventory with contrasting articulatory features.

Features-values To ensure compatibility with external resources, we linked our features and values to other standards. All datasets contain mappings to UniMorph (Sylak-Glassman et al., 2015) and Universal Dependencies (Nivre et al., 2016) dialects. Additionnally, *ParaLiv* maps to the referential used by the Livonian Institute and the University of Tartu dialect corpus (Lindström et al., 2022), *ParaFin* maps to the Omorfi encoding and *ParaKar* to the VepKar unique identifiers. These mappings have proven valuable in extracting token frequencies (see below). An overview of the mappings offered in the three datasets is provided for the illative singular cell in Table 3.

Overabundance and defectivity In the Paralex format, each wordform is assigned a record. If two forms are available for a given cell, a case of overabundance (Thornton, 2019), two records are created. If a cell has no known form, a record is still created with the label `#DEF#`. For such non-canonical phenomena, we provide semantic annotations to distinguish overabundant forms and to make explicit the reason for defectivity. For instance, in Finnish, the third person possessive suffix takes two forms: *-nsA* or *-Vn*. Such forms are tagged `poss_nsA` and `poss_Vn`. A record can have several tags.[5] Concerning defectivity, Omorfi and VepKar tend to provide extensive paradigms.[6] This can partly be explained by the difficulty of assessing the defectivity of a given form, due to low frequency effects (Nikolaev and Bermel, 2023).

Frequencies Paralex lexicons can optionally store frequencies at three different levels: cells, forms and lexemes. As for our lexicons, we provide all frequencies for Finnish and Livvi, but only cell frequencies for Livonian.

The frequencies were extracted from the Finnish dataset provided with the LASTU software (Itkonen et al., 2024), which in turn relies on the Finnish Parsebank (Luotolahti et al., 2015). We used the frequency table for forms occuring at least 10 times in the parsebank. We matched the universal dependency features used in the original dataset with our own cells and ruled out all inconsistent annotations. In further versions of the dataset, we plan to introduce frequencies directly extracted from the parsebank. For lexemes, we use the cumulated lexeme frequencies already provided by the LASTU dataset. For Karelian, we used the annotated VepKar corpus to extract form, lexeme and cell frequencies. For Livonian, we extracted word frequencies from the Estonian Dialects Corpus (Lindström et al., 2019, 2022) and grouped them by cell. This corpus was to small to assign a frequency to the lexemes or to the forms.

4 Conclusion

We introduced inflected lexicons for three Finnic languages: Livonian, Finnish and Livvi. We reviewed current practices in Uralic language resources and emphasized the importance of rich, machine-readable formats to facilitate cross-linguistic studies of morphological systems. We presented the design choices for our datasets and introduced our linguistically motivated grapheme-to-phoneme rules. We outlined the annotations that we performed. Appendix A showcases the main tables of one of the resulting datasets.

Although we did our best to manually check the transcriptions by evaluating random samples of forms and by carrying out targeted verifications, it is very likely that some mistakes remain, especially for loanwords. In addition to improved transcriptions, further versions should include more morphological annotations (e.g. information on stem gradation according to traditional descriptions) and reference other sources of frequencies (especially for Finnish). The datasets could also be extended to verbal inflection.

[5] Some forms follow idiosyncratic patterns and are not tagged. The percentage of untagged forms is: 1.27% in *ParaFin*, 4.24% in *ParaLiv* and 5.08% in *ParaKar*.

[6] In Omorfi, only *pluralia tantum* appear as defective.

Acknowledgments

We are thankful to the Livonian Institute, the Karelian Research Center and the developers of Omorfi for providing access to their morphological resources. The help of Tuuli Tuisk and Valts Erntreits was precious in answering our questions about Livonian phonology and the resources of the Livonian Institute. Olivier Bonami provided comments on this paper and Sacha Beniamine answered questions related to the datasets.

References

Anneli Arhimaa. 2022. Karelian. In Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik, editors, *The Oxford Guide to the Uralic Languages*, pages 269–290. Oxford University Press, Oxford.

R. Harald Baayen, Yu-Ying Chuang, Elnaz Shafaei-Bajestan, and James P. Blevins. 2019. The Discriminative Lexicon: A Unified Computational Model for the Lexicon and Lexical Processing in Comprehension and Production Grounded Not in (De)Composition but in Linear Discriminative Learning. *Complexity*, 2019:39.

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik. 2022. *The Oxford Guide to the Uralic Languages*. Oxford Guides to the World's Languages. Oxford University Press, Oxford.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, pages 840–855, Marseille, France. European Language Resources Association (ELRA).

Sacha Beniamine, Mari Aigro, Matthew Baerman, Jules Bouton, and Maria Copot. 2024. Eesthetic: A Paralex Lexicon of Estonian Paradigms. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5526–5537, Torino, Italia. European Language Resources Association (ELRA) and International Committee on Computational Linguistics (ICCL).

Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. Paralex: a DeAR standard for rich lexicons of inflected forms. In *International Symposium of Morphology (ISMo 2023)*, Nancy, France. https://www.paralex-standard.org.

Sacha Beniamine, Olivier Bonami, and Ana R. Luís. 2021. The fine implicative structure of European Portuguese conjugation. *Isogloss. Open Journal of Romance Linguistics*, 7:1–35.

Robert Bielecki. 2009. On the Nature of the Accusative in Finnish. *Lingua Posnaniensis*, 51(1):19–38.

Tatyana Boyko, Nina Zaitseva, Natalia Krizhanovskaya, Andrew Krizhanovsky, Irina Novak, Nataliya Pellinen, and Aleksandra Rodionova. 2022. The Open Corpus of the Veps and Karelian Languages: Overview and Applications. *KnE Social Sciences*, 7(3):29–40.

Valts Ernštreits, Tiit-Rein Viitso, and Milda Kurpniece. 2024. Livonian morphology database. http://www.livonian.tech.

Dan Fowler, Jo Barratt, and Paul Walsh. 2017. Frictionless Data: Making Research Data Quality Visible. *International Journal of Digital Curation*, 12(2):274–285.

Anders Holmberg and Urpo Nikanne, editors. 1993. *Case and Other Functional Categories in Finnish Syntax*. Number 39 in Studies in Generative Grammar. De Gruyter Mouton.

Sami Itkonen, Tuomo Häikiö, Seppo Vainio, and Minna Lehtonen. 2024. LASTU: A psycholinguistic search tool for Finnish lexical stimuli. *Behavior Research Methods*, 56(6):6165–6178.

Fred Karlsson. 1983. *Suomen kielen äänne- ja muotorakenne*. Werner Söderström, Porvoo, Finland.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J. Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Christo Kirov, John Sylak-Glassman, Roger Que, and David Yarowsky. 2016. Very-large Scale Parsing and Normalization of Wiktionary Morphological Paradigms. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 3121–3126, Portorož, Slovenia. European Language Resources Association (ELRA).

Natalia Krizhanovskaya, Irina Novak, and Nataliya Pellinen. 2024. Pravila generacii imennyh slovoform po minimizirovannomu šablonu dlâ novopis'mennyh variantov sobstvenno karel'skogo i livvikovskogo narečij.

Liina Lindström, Pärtel Lippus, and Tuuli Tuisk. 2019. The online database of the University of Tartu Archives of Estonian Dialects and Kindred Languages and the Corpus of Estonian Dialects. *Uralica Helsingiensia*, (14):327–350.

Liina Lindström, Triin Todesk, and Maarja-Liisa Pilvik. 2022. Corpus of Estonian Dialects.

Juhani Luotolahti, Jenna Kanerva, Veronika Laippala, Sampo Pyysalo, and Filip Ginter. 2015. Towards Universal Web Parsebanks. In *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 211–220, Uppsala, Sweden. Uppsala University.

Robert Malouf. 2017. Abstractive morphological learning with a recurrent neural network. *Morphology*, 27(4):431–458.

Robert Malouf, Farrell Ackerman, and Arturs Semenuks. 2020. Lexical databases for computational analyses: A linguistic perspective. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 446–456, New York, New York. Association for Computational Linguistics (ACL).

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, pages 3922–3931, Marseille, France. European Language Resources Association (ELRA).

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for Many Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2711–2714, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexandre Nikolaev and Neil Bermel. 2023. Studying negative evidence in Finnish language corpora. *Word Structure*, 16(2-3):206–232.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Irina Novak, Natalia Krizhanovskaya, Tat'jana Bojko, and Nataliya Pellinen. 2020. Development of rules of generation of nominal word forms for new-written variants of the Karelian language. *Bulletin of Ugric studies*, 10(4):679–691.

Irina Novak, Martti Penttonen, Aleksi Ruuskanen, and Lea Siilin. 2022. *Karelian in Grammars : A study of phonetic and morphological variation*. Karelian Research Centre of the Russian Academy of Sciences, Petroskoi.

Flammie Pirinen, Sjur Moshagen, and Katri Hiovain-Asikainen. 2023. GiellaLT — a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 643–649, Tórshavn, Faroe Islands. University of Tartu Library.

Tommi A. Pirinen. 2015. Development and Use of Computational Morphology of Finnish in the Open Source and Open Science Era: Notes on Experiences with Omorfi Development. *SKY Journal of Linguistics*, 28:381–393.

Tommi A. Pirinen, Inari Listenmaa, Ryan Johnson, Francis M. Tyers, and Juha Kuokkala. 2017. Open morphology of Finnish. University of Helsinki.

Lauri Posti. 1942. *Grundzüge der livischen Lautgeschichte*. Number 75 in Mémoires de la Société Finno-Ougrienne. University of Helsinki, Helsinki.

Raija Pyöli. 2011. *Livvinkarjalan kielioppi*. Karjalan kielen seura, Helsinki.

Kari Suomi, Juhani Toivanen, and Riikka Ylitalo. 2008. *Finnish sound structure: phonetics, phonology, phonotactics and prosody*. Number 9 in Studia Humaniora Ouluensia. University of Oulu, Oulu.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A Language-Independent Feature Schema for Inflectional Morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 674–680, Beijing, China. Association for Computational Linguistics (ACL).

Anna Maria Thornton. 2019. Overabundance: A Canonical Typology. In Franz Rainer, Francesco Gardani, Wolfgang U. Dressler, and Hans Christian Luschützky, editors, *Competition in Inflection and Word-Formation*, number 5 in Studies in Morphology, pages 223–258. Springer, Cham.

Tuuli Tuisk. 2015. Acoustics of Stød in Livonian. In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK. University of Glasgow.

Tuuli Tuisk. 2016. Main features of the Livonian sound system and pronunciation. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics*, 7(1):121–143.

VepKar. 2009/2024. Open Corpus of Veps and Karelian languages. http://dictorpus.krc.karelia.ru/.

Tiit-Rein Viitso. 2007. Livonian Gradation : Types and Genesis. *Linguistica Uralica*, 43(1):45.

Tiit-Rein Viitso and Valts Ernštreits. 2012. *Līvõkīel-ēstikīel-leţkīel sõnārõntõz = Liivi-eesti-läti sõnaraamat = Lībiešu-igauņu-latviešu vārdnīca*. University of Tartu, Tartu.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018.

A Appendix: Sample tables from the *ParaLiv* dataset

(a) The forms table

form_id	lexeme	cell	orth_form	phon_form	analysed_orth_form	analysed_phon_form	defectiveness_tag	overabundance_tag
vȭrbõz_22999-dat.pl	vȭrbõz_22999	dat.pl	vȭrbõdõn	v ɯː r b ɯ d ɯ n	vȭrbõdõn	v ɯː r b ɯ d ɯ n		
irī_13393-ill.sg-1	irī_13393	ill.sg	irī	i r iː	irī	i r iː		illsg_without_z
ēřškõmōŗapȭzõ_12800-ine.sg	ēřškõmōŗapȭzõ_12800	ine.sg	ēřškõmǭŗapȭzõs	eː r ʃ k ɯ m ɔː rʲ ɑ pʼ ɯː z ɯ s	ēřškõ|mǭŗa|pȭzõs	eː r ʃ k ɯ + m ɔː rʲ ɑ + pʼ ɯː z ɯ s		
passõr_18233-dat.pl	passõr_18233	dat.pl	passõrdõn	p ɑ sː ɯ r d ɯ n	passõrdõn	p ɑ sː ɯ r d ɯ n		
vikāt_22643-ine.pl	vikāt_22643	ine.pl	vikātis	v i kʼ ɑː tʼ i s	vikātis	v i kʼ ɑː tʼ i s		
nȭŗkõz_17668-ine.pl	nȭŗkõz_17668	ine.pl	nȭŗkõžis	n ɯː rʲ k ɯ ʒ i s	nȭŗkõžis	n ɯː rʲ k ɯ ʒ i s		
sīlmadkōp_20193-inm.pl	sīlmadkōp_20193	inm.pl	sīlmadkōpõdõks	s iː l m ɑ d k oː pʼ ɯ d ɯ k s	sīlmad|kōpõdõks	s iː l m ɑ d + k oː pʼ ɯ d ɯ k s		
kȭrtami_15156-gen.pl	kȭrtami_15156	gen.pl	kȭrtamizt	k ɯː r t ɑ m i s t	kȭrtamizt	k ɯː r t ɑ m i s t		
suoļm_20758-nom.pl	suoļm_20758	nom.pl	sūoļmõd	s uːo lʲ m ɯ d	sūoļmõd	s uːo lʲ m ɯ d		
salāndõm_19842-ela.pl	salāndõm_19842	ela.pl	salāndõmist	s ɑ l ɑː n d ɯ m i s t	salāndõmist	s ɑ l ɑː n d ɯ m i s t		

(b) The lexemes table

lexeme_id	label	inflection_class	POS
armāiga_11913	armāiga	101	noun
skūolsoppiji_20388	skūolsoppiji	286	noun
ministrij_16883	ministrij	199	noun
kūjabulā_15356	kūjabulā	83	noun
pūlēḑ_19079	pūlēḑ	234	noun
azūmsõnā_11992	azūmsõnā	83	noun

(c) The cells table

cell_id	POS	unimorph	ud	livonian_tech	tartu	frequency
nom.sg	noun	N;NOM;SG	Case=Nom|Number=Sing	NomSg	sg.nom.	5410
gen.sg	noun	N;GEN;SG	Case=Gen|Number=Sing	GenSg	sg.gen.	3941
dat.sg	noun	N;DAT;SG	Case=Dat|Number=Sing	DatSg	sg.dat.	762
prt.sg	noun	N;PRT;SG	Case=Prt|Number=Sing	PrtSg	sg.part.	1890
inm.sg	noun	N;INS;SG	Case=Ins|Number=Sing	InmSg	sg.tr.	177
ill.sg	noun	N;IN+ALL;SG	Case=Ill|Number=Sing	IllSg	sg.ill.	877

(d) The sounds table

sound_id	CLTS_id	syllabic	stress	long	half-long	consonantal	…
jː	jː	0		1		0	…
kʼ	kʼ	0		0	1	1	…
iˀu		1	1	0		0	…
rʲʼ		0		0	1	1	…
ieuː		1	1	0		0	…
oːˀ		1	1	1		0	…

(e) The tags table

tag_id	tag_column_name	comment
defective	defectiveness_tag	defective for unknown reasons
pluralia_tantum	defectiveness_tag	defective in singular because pluralia tantum
illsg_without_z	overabundance_tag	a parallel form for illatives; without z final consonant
illsg_with_z	overabundance_tag	a parallel form for illatives; with z final consonant
elasg_with_õ	overabundance_tag	a parallel form for consonantal words; with õ final vowel
elasg_without_õ	overabundance_tag	a parallel form for consonantal words; without õ final vowel

(f) The graphemes table

grapheme_id	comment	canonical_order
a		3
ā		4
k		18
o		24
ŗ		29
z		39

(g) The values table

value_id	label	POS	feature	unimorph	ud	livonian_tech	tartu	canonical_order
nom	nominative	noun	case	NOM	Case=Nom	Nom	nom	1
gen	genitive	noun	case	GEN	Case=Gen	Gen	gen	2
prt	partitive	noun	case	PRT	Case=Par	Prt	part	3
dat	dative	noun	case	DAT	Case=Dat	Dat	dat	4
inm	instrumental-comitative	noun	case	INS	Case=Ins	Inm	tr	5
ill	illative	noun	case	IN+ALL	Case=Ill	Ill	ill	6

Table 4: Excerpts from the forms, lexemes, cells, sounds, tags, graphemes, features tables from the *ParaLiv* package. Primary keys have a grey shading.