Data release: Digitized Dialect Atlas of Finnish by Lauri Kettunen

Jenni Santaharju[a,b], Kaj Syrjänen[c, d], Terhi Honkola[c], Perttu Seppä[b], Outi Vesakoski[a] and Unni Leino[d]

[a] *University of Turku, School of Languages and Translation Studies*
[b] *University of Helsinki, Organismal and Evolutionary Biology Research Programme, Faculty of Biological and Environmental Sciences*
[c] *University of Turku, Department of Biology*
[d] *Tampere University, Faculty of Information Technology and Communication Sciences*

Abstract

In this paper, we present a new digitized version of the Dialect Atlas of Finnish and offer the data as open access. The Atlas was compiled in the 1920s and 1930s to document the variation of Finnish language. It provides the only description of Finland's historical linguistic landscape before urbanization and media began to homogenize the dialectal differences. The Atlas was digitized at the turn of the millennium, and a corrected version was published online by the Institute for the Languages of Finland ten years ago. Here we provide an improved version of the Atlas in a more user-friendly format, along with description of the linguistic traits. Additionally, we supplement the Atlas with polygons indicating the borders of historical municipalities (study units), enabling spatial analyses and visualization. The selected data formats are the same that were used in recent dialectometric studies, which we briefly describe. This data release will improve accessibility and reusability of the Atlas.

Keywords

Dialectometry, Finnish dialects, Linguistic landscape, Language evolution, Spatial analyses

1. Background

In this data release paper, we present a new digitized version of the Dialect Atlas of Finnish (1940a) to promote its use in both linguistic and interdisciplinary research. The Atlas was compiled by Lauri Kettunen (1885–1963) in the 1920s and 1930s as a part of an international trend of collecting the variation of local languages into atlases. He travelled through the Finnish-speaking municipalities – some of which are nowadays located in Sweden, Norway and Russia – and interviewed people that would have preserved the "traditional" dialectal variation (Kettunen 1930a,b; 1940a,b; more precise description of the data collection in Vesakoski et al. 2024). The Atlas represents the most comprehensive data of the dialectal variation in Finnish and is the only data available of its historical linguistic landscape. The Atlas thus allows for studying the "natural" linguistic landscape of Finnish; the linguistic landscape became more mixed after urbanization, utilization of mass media and mass movement of Karelian evacuees to the current Finland area during the Second World War (e.g. Lynch et al. 2022).

Until recently, the use of dialect atlases has been limited by the absence of methods to integrate the information of individual isoglosses (the distribution ranges of the dialectal traits). The digital Atlas in numerical form aims to address this gap. The work by BEDLAN team (Biological Evolution and the Diversification of Languages, www.bedlan.net), has been part of the new, computational paradigm in geographical dialectology that aims to tame the "uncontrollable chaos of the linguistic atlases" (Goebl 2010). The data frames offered here are built for such purposes.

Digital Humanities in the Nordic and Baltic Countries 2024: From Experimentation to Experience, May 27-31, 2024, Reykjavík, Iceland

EMAIL: jenni.santaharju@utu.fi (A. 1); kaj.syrjanen@utu.fi (A. 2); terhi.honkola@utu.fi (A. 3); perttu.seppa@helsinki.fi (A. 4); outi.vesakoski@utu.fi (A. 5); unni.leino@tuni.fi (A. 6)
ORCID: 0000-0002-7925-2715 (A. 1); 0000-0002-9795-7284 (A. 2); 0000-0002-3330-0857 (A. 3); 0000-0001-5393-6943 (A. 4); 0000-0002-7220-3347 (A. 5); 0000-0003-3917-0026 (A. 6)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

2. The data

Here we provide an easy-to-use and well-documented raw data of the Dialect Atlas of Finnish. It includes 213 linguistic traits organised into 213 maps, each of which describes the spatial distribution of the variants of one linguistic trait across 525 municipalities. Thus, the study unit is a municipality (or parish). Each linguistic trait has 2–14 variants, and most municipalities have only one of these variants, but sometimes two, and in very rare cases three or four. Most of the traits describe morpho-phonological variation of the dialects: for example, how consonant gradation varies across dialects (such as the case of *pata* : *padan* / *paðan* / *paran* / *palan* / *poan* etc.) or whether a schwa-vowel exists between certain consonants (for example 'old' expressed as *vanha* vs. *van$^{(a)}$ha* vs. *vanaha).* Atlas provides no sociolinguistic information of the people interviewed.

The first version of the digitized data was prepared by Embleton & Wheeler (1997, 2000) at the University of Toronto. It was further curated by the BEDLAN team and published as an undocumented version in Finnish in collaboration with the Institute for the Languages of Finland (see http://urn.fi/urn:nbn:fi:csc-kata20151130145346403821). This data format is rather non-transparent with no metadata about the actual linguistic traits or background of the collections.

To facilitate the use of the digitized Atlas, we offer a complete release of the data, including English translations of the linguistic traits. To enable the spatial analysis and visualization of the dialectal data, we provide geospatial polygon information that describes the boundaries and central coordinates of the historical municipalities used in the Atlas. All this makes this valuable dialect data more findable, accessible, interoperable and reusable (FAIR Principles). Finally, we also provide 26 dialectal variants that were missing from the earlier version of the digitized Atlas.

The new digitized version of the Dialect Atlas of Finnish is based on coding practices necessitated in the statistical analyses used in Syrjänen et al. (2016) and in Santaharju et al. (resubmitted revision). Syrjänen et al. (2016) coded the data so that each municipality had a maximum of two linguistic variants in a given linguistic trait. Instead, Santaharju et al. (resubmitted revision) binarised the multistate data, giving each linguistic variant across the whole data a presence/absence state so that each municipality could be associated with multiple variants. Similar cognate coding practice has been used also in phylolinguistic studies. A more detailed data description and the digitized Atlas data are available in Zenodo (Santaharju et al. 2024). Scanned maps of the original Atlas are available in (http://kettunen.free.nf) and will be added to URHIA (Uralic Historical Atlas, www.urhia.fi; see Roose et al. this volume).

3. Data applications so far

The Dialect Atlas of Finnish has served as a valuable resource in Finnish dialectology for both teaching and research (Syrjänen et al. 2016; Aarikka 2023; Vesakoski et al. 2024 and the references therein). The data format presented here enables traditional dialectological studies as well as computational approaches. So far, the data has been utilized in clustering of the Finnish dialects (Syrjänen et al. 2016) and in studying the drivers of dialectal differentiation (Honkola et al. 2018) and convergence (Santaharju et al. resubmitted revision). All these studies built on a population genetic toolkit, allowing for the merging of isoglosses and the formation of dialect clusters with fuzzy borders, and producing new measures of dialectal variation for municipalities and regional dialects. The dialect clusters with fuzzy borders have been produced by using the Bayesian model-based clustering methods, which can divide the data to the chosen or optimal number of dialect clusters and estimate an affinity for each municipality to each dialect cluster (see Fig. 1), enabling the further measures.

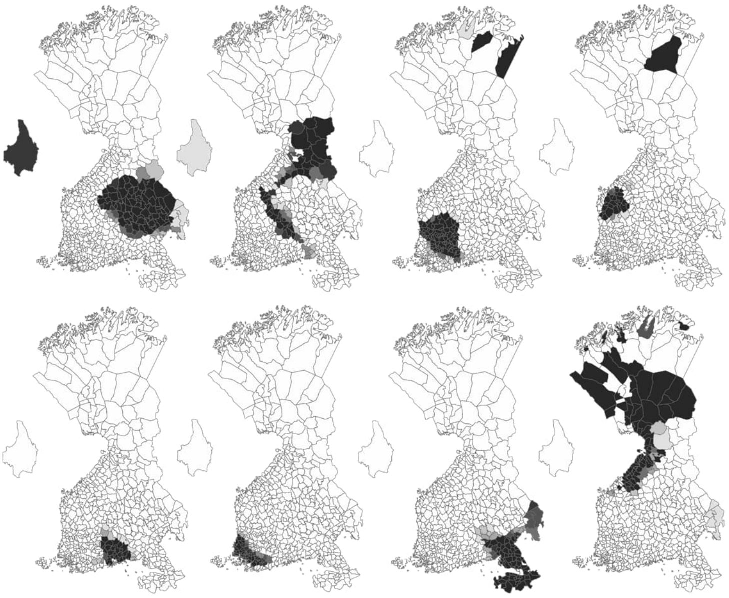

Figure 1: An example of forming the regional dialects based on the Dialect Atlas of Finnish. Here Bayesian clustering analysis, implemented in software BAPS (Corander et al. 2003), was used to divide the data into eight dialect clusters (K=8), simultaneously obtaining affinities of municipalities to each cluster. The darker the municipality, the greater its dialect affinity to the particular dialect cluster. Some municipalities are located outside the area today known as Finland, in Sweden, Norway, and Russia. Municipality of Värmland on the top left is not in its real geographical location in southwestern Sweden.

The new measures based on affinities of each municipality to each dialectal area (Syrjänen et al. 2016), make the historical municipalities interoperable between various types of data, such as genetic and cultural data. This opens up a possibility to cross-analyse multidisciplinary municipality-wise data to reconstruct an interdisciplinary view to the human past in Finland. For example, Lynch et al. (2022) estimated the linguistic differentiation between municipalities and studied if Karelian evacuees during the Second World War tended to stay in the area of arrival based on the linguistic (versus ecological) similarity and betterment of the area. In contrast, Nitch et al. (in press) used the dialectal clusters to study the areal spread of communicative diseases in historical Finland. We hope that this data release of the digitized Dialect Atlas of Finnish by Lauri Kettunen will serve as a recognition of the enduring value of this meticulous work from 100 years ago. This data release is part of the Uralic Trove digital data infrastructure (Vesakoski et al. this volume).

4. Acknowledgements

We thank Kotus for providing the data, Toni Suutari for collaboration, Jyri Lehtinen and Kaj Syrjänen for data refinement, Peter Cady and Michael Dunn for proposals to improve the data, and Lotta Aarikka and Meeli Roose for commenting the manuscript. The work was funded by the Kone Foundation (“BEDLAN”, “SumuraSyyni”, “AikaSyyni”, “Kippo”), the Research Council of Finland (grant no. 352727), the Otto A. Malm foundation and the Ella and Georg Ehrnrooth Foundation.

5. References

Aarikka, Lotta. 2023. “Dialect and its study. Perspectives on the history and language ideologies of Finnish dialectological research 1871–2017.” *Annales universitatis turkuensis.* https://www.utupub.fi/handle/10024/174906

Corander, Jukka, Patrik Waldmann and Mikko J. Sillanpaa. 2003. “Bayesian analysis of genetic differentiation between populations.” Genetics 163.367–74. https://doi.org/10.1093/genetics/163.1.367.

Embleton, Sheila and Eric Wheeler, S. 1997. “Finnish dialect atlas for quantitative studies.” Journal of *Quantitative Linguistics* 4.99-102. https://doi.org/10.1080/09296179708590082.

Embleton, Sheila, M. and Eric Wheeler, S. 2000. "Computerized dialect atlas of Finnish: Dealing with ambiguity." *Journal of Quantitative Linguistics* 7.227-31. https://doi.org/10.1076/jqul.7.3.227.4109.

Goebl, Hans. 2010. "Dialectometry and quantitative mapping." *Language and Space. An International Handbook of Linguistic Variation*. Berlin: De Gruyter. https://doi.org/10.1515/9783110219166.1.433.

Honkola, Terhi, Kalle Ruokolainen, Kaj J. J. Syrjänen, Leino Unni-Päivä, Ilpo Tammi, Niklas Wahlberg and Outi Vesakoski. 2018. "Evolution within a language: Environmental differences contribute to divergence of dialect groups." *BMC Evolutionary Biology* 18. https://doi.org/10.1186/s12862-018-1238-6.

Kettunen, Lauri. 1930a. "Suomen murteet I. Murrenäytteitä." Helsinki: *Suomalaisen Kirjallisuuden Seuran Kirjapainon osakeyhtiö.*

Kettunen, Lauri. 1930b. "Suomen Murteet II. Murrealueet." Helsinki: *Suomalaisen Kirjallisuuden Seura.*

Kettunen, Lauri. 1940. "Suomen Murteet III A. Murrekartasto." Helsinki: *Suomalaisen Kirjallisuuden Seura*.

Kettunen, Lauri. 1940b. "Suomen Murteet III B. Selityksiä Murrekartastoon." Helsinki: *Suomalaisen Kirjallisuuden Seura.*

Lynch, Robert, John Loehr, Virpi Lummaa, Terhi Honkola, Jenni Pettay and Outi Vesakoski. 2021. Socio-cultural similarity with host population rather than ecological similarity predicts success and failure of human migrations. Proceedings of the Royal Society B 289.20212298. http://doi.org/10.1098/rspb.2021.2298.

Nitch, Aïda, Virpi Lummaa, Tarmo Ketola, Terhi Honkola, Outi Vesakoski and Michael Briga. "Is dialect clustering relevant for spatial epidemiology? A test on infectious diseases epidemics in historical Finland." In press in iScience.

Roose, Meeli, Tua Nylén, Petro Pesonen, Harri Tolvanen and Outi Vesakoski. "Uralic Historical Atlas (URHIA): Interactive Web App for Spatial Data." *Digital Humanities in the Nordic and Baltic Countries Publications.* This volume.

Santaharju, Jenni, Terhi Honkola, Perttu Seppä, Kaj Syrjänen, Unni Leino and Outi Vesakoski. "Linguistic convergence and its drivers in Finnish dialects". Resubmitted revision.

Santaharju, Jenni, Kaj Syrjänen, Terhi Honkola, Perttu, Seppä, Outi Vesakoski and Unni Leino (2024): New version of the digitized Dialect Atlas of Finnish by Lauri Kettunen. Zenodo. https://doi.org/10.5281/zenodo.10078078.

Syrjänen, Kaj, Terhi Honkola, Jyri Lehtinen, Antti Leino and Outi Vesakoski. 2016. "Applying population genetic approaches within languages: Finnish dialects as linguistic populations." *Language Dynamics and Change* 6.235 – 83. https://doi.org/10.1163/22105832-00602002.

Vesakoski, Outi; Lotta Aarikka and Jenni Santaharju. 2024. "Tieteen matkamiehen siivellä – Lauri Kettusen Suomen murteet tutkimusaineistona ennen ja nyt." *Saarte keeled. Ellen Niidi juubeliraamat.* Tallin*:* Eesti Teaduste Akadeemia Emakeele Seltsi. 83. 142–165.

Vesakoski, Outi, Michael Dunn, Meeli Roose and Jenni Santaharju. "The Uralic Trove (UraLaari) – The digital data infrastructure of speaker areas of Uralic languages" *Digital Humanities in the Nordic and Baltic Countries Publications.* This volume.