ANNELI SARHIMAA
University of Mainz

Finnic data sets in the ELDIAdata databank

1. Introduction

The ELDIAdata is a digital databank containing all empirical materials collected during the EU FP-7 research project ELDIA (European Language Diversity for All). ELDIA was an international, interdisciplinary project coordinated by the University of Mainz, Germany, and conducted by experts in applied linguistics and sociolinguistics, law, social studies, and statistics in 2010–2013. The overarching aim was to contribute to a better understanding of how local, national, and international (vehicular) languages interact in contemporary Europe, and to enhance the reconceptualisation, re-evaluation, and promotion of individual and societal multilingualism.

The empirical data were collected in eight countries from speakers of thirteen Finno-Ugric minority languages, ten of which belong to the Finnic group: Meänkieli, Kven, Finnish in Sweden, Estonian in Finland, Estonian in Germany, Karelian in Finland, Karelian in Russia, Veps, Võro, and Seto. In addition to the primary study populations, that is, the minority groups whose languages were at stake, a control group representing all other citizens of the countries where the investigated minorities live was surveyed and interviewed, as well.

The most concrete goal of ELDIA was to create a European Language Vitality Barometer (EuLaViBar). The barometer is a tool for measuring the level of language vitality and it helps identify areas within which the assessed language requires more societal support. At the most general level, it serves as a testable model for assessing not only the vitality but also the social significance of different languages in multilingual

Multilingual Finnic. Language contact and change. 367–377.
Uralica Helsingiensia 14. Helsinki 2019.
‹https://doi.org/10.33341/uh.85042›

communities. The barometer also can be applied to other languages beyond those in the Finno-Ugric family and those spoken in Europe.

The results of ELDIA have been published in case-specific reports, in an overview report (Laakso et al. 2013), and as a toolkit supervising end users in creating and using the barometer (Spiliopoulou Åkermark et al. 2013). The reports and the toolkit are available online on the project website. A book placing the project results in wider academic and language-political contexts appeared in 2016 (Laakso et al. 2016).

2. An overview of the Finnic data sets

The ELDIAdata databank contains all the statistical data collected with a large-scale questionnaire survey as well as all interviews carried out with speakers of the investigated minority languages and with representatives of the control groups. The databank consists of two major parts: the minority language target group database and the control group database. Table 1 summarizes the Finnic data sets.

As Table 1 indicates, joint control group survey questionnaires were used for minorities that were studied in the same country, that is, for Meänkieli and Finnish in Sweden, for North Saami and Kven in Norway, for Karelian and Estonian in Finland, for Karelian and Veps in Russia, and for Seto and Võro in Estonia. The data sets from Sweden and Germany are not complete. The case study concerning Finnish in Sweden had to be given up due to organisational and other problems in 2011, and so only survey data were collected there. In Germany, contrasting a couple of thousand Estonians who live scattered all over the country with the 82 million other inhabitants would have been utterly senseless, and so no control group data were collected there.

3. How the data were collected

The empirical data collection in ELDIA aimed at accumulating new information on the investigated Finno-Ugric minority languages in a systematic manner for the purposes of developing the barometer. Another aim was to fill gaps in the existing research with the help of a multilingual corpus containing statistical and interview data.

Case study	Number of questionnaires distributed		Number of questionnaires returned		Response rate in %		Number of interviews		Interview material (hours:minutes)	
	Target group	Control group	Target group*	Control group*	Target group	Control group	Indiv.	Group	Indiv.	Group
Meänkieli in Sweden	941	1 000	554	227	58.87	22.7	7	8	07:15	13:11
Finnish in Sweden	1 000		369		36.9		–	–	–	–
Kven in Norway	1 500	1 000	85	107	5.67	10.7	8	8	04:45	10:10
Veps in Russia	301	302	299	302	99.34	100	7	6	05:51	05:31
Karelian in Russia	301		296		98.34		6	6	04:16	07:17
Karelian in Finland	1 034	800	356	146	34.43	18.25	8	8	08:33	12:13
Estonian in Finland	800		170		21.25		8	8	09:58	11:18
Estonian in Germany	420	none	71	–	16.9	–	8	3	13:03	05:33
Seto in Estonia	418	1 000	294	363	70.33	36.3	8	8	06:38	06:20
Võro in Estonia	409		296		72.37		8	8	05:54	07:56
In total	**7 124**	**4 102**	**2 790**	**1 145**	**51.44**	**37.59**	**68**	**63**	**65:12**	**80:04**

Table 1. The Finnic data sets within the ELDIAdata databank

3.1. Data collection tools

The new data were collected using the following tools:

- a unified survey questionnaire designed for the minorities;
- a unified survey questionnaire designed for the control groups;
- a semi-structured matrix for focus group interviews with minority stakeholder and speaker groups;
- a semi-structured matrix for focus group interviews with selected representatives of the control groups consisting of politicians, authorities, and representatives of media;
- a semi-structured matrix customised to be case-specific for interviews with individual speakers of the minority language at issue.

The questionnaire survey among the investigated minority communities served as the main means for collecting new information on the current use and the current state of the investigated minority languages. The survey sought to provide a broad and general insight into the state and the use of the investigated languages in a way that also would facilitate generalizable comparisons between the minority groups.

The minority language questionnaire was structured around the following main topics:

Part A. Personal background information on the respondent
Part B. Background information on the respondent's language use
Part C. The respondent's language skills (minority language, majority language, other languages)
Part D. The respondent's use of minority, majority, and other language(s) in different domains
Part E. Language attitudes and the respondent's desire to use languages
Part F. Public language use vs. private language use
Part G. Consumption and active use of languages by the respondent in different media

There were 63 questions in the minority questionnaire, but as there was a high number of sub-questions, the total number of questions was well over 300. The statistical datasets included in the databank contain only variables derived from the closed questions of the survey questionnaire; the variables total exactly 340.

The control group survey contained the same primary topics as the minority language survey, with the exception that Part F "Public language use vs. private use" was omitted. The contents of the control group questionnaire were largely the same as those of the minority group questionnaire; however, some questions not relevant for majority language speakers were omitted and some questions were formulated differently. The control group survey questionnaire included 47 questions. Again, due to the large number of sub-questions, the variables in the control group statistical data set total 280.

The ELDIA interview design aimed primarily at completing the survey data with in-depth qualitative insights into the themes covered by the survey. The interviews also offered the researchers the possibility of collecting information that would shed some new light on the case-specific

research gaps that had been identified in ELDIA desk research. The original plan was to conduct eight individual interviews with members of the minority, eight focus group interviews with members of the minority, and three focus group interviews with representatives of society at large.

The eight individual interviews and the eight focus group interviews with members of the minorities were carried out in five age groups (18–29, 30–49, 50–65, and over 65). In order to obtain a wider perspective on the current parental generation, the 30–49 age group was divided into two groups with a group of women and a group of men interviewed separately. According to the project design, the three minority focus groups should have covered (i) minority politicians and civil servants belonging to the investigated minority group, (ii) minority activists and (iii) representatives of the minority media. However, in several ELDIA case studies it only was possible to create a separate group composed of activists while the other two groups had to be combined. In some case studies it was only possible to find people for the age-based focus groups while the other focus group interviews had to be omitted completely.

The original data collection design included two focus groups involving representatives of the control group: one for politicians and civil servants dealing with (minority-)language matters, and one for representatives of the majority-language media. The case study focusing on Estonian in Germany did not involve any control group, therefore, these interviews were not conducted. In some case studies such as Hungarian in Austria and for Kven and North Saami in Norway, there were problems in getting enough interviewees, especially for the focus group interview with politicians and civil servants.

All focus group interviews followed a joint thematic interview template that was modified to meet the case-specific needs. The main themes were the interviewee's bilingualism or multilingualism, their use of different languages in everyday life and views on bilingualism or multilingualism as being either an asset or a problem. The interviewees were also asked about their perception of the term 'minority' and their opinion of the role of language in an individual's identity; for example, whether knowing the minority language is necessary for identification as a member of a given minority. Yet another issue was the perception of the investigated minority and its language by the society, the role of the investigated minority language in the society at

issue as well as the role of societal diversity in general. Special attention was also drawn to mapping the interviewees' opinions concerning the responsibility of society in maintaining and revitalising the investigated minority language and on their views concerning the future of the minority language in a ten-year perspective.

The individual interviews were conducted with one male and one female per age group. The interviews were designed to collect in-depth information on four thematic fields: mother tongue; other languages; attitudes towards multilingualism; and languages and modernisation. The interview format was semi-structured in the sense that the topics were predefined but it was left to the interviewer to formulate the questions in a way that will be experienced as maximally "natural" in the interview situation and to ask further questions as suitable in the communication situation.

3.2. Principles for identifying survey respondents and interviewees

As explained earlier, concurrent sample surveys were conducted in eight countries to obtain information on the target populations (minority language groups) and the corresponding majority populations (control groups) in all other countries except Germany. Originally, the idea was to obtain about 300 responses from each survey population. As the non-response rates are high in mail surveys, the sample size had to be inflated by the anticipated non-response. Not even these measures were always sufficient: as can be read in Table 1, in some cases the final number of respondents remained clearly below the target.

Ideally, the sampling design should have been a true random sample in each country and study population. However, the project team was aware that in some countries there are no proper ways to carry out real probability sampling methods among the study populations. Those problems arise either from the lack of availability of comprehensive sampling frames like population registers, or legislation on disclosure control. All researchers involved in sample selection and data collection co-operated closely with the project statistician Kari Djerf (University of Helsinki) and with Karl Pajusalu (University of Tartu) who was in charge for the fieldwork in ELDIA. The best possible solution was sought for each case

study to assure the maximal comparability of the obtained data. Ultimately, random sampling could be applied in most cases in one form or another. Yet, in the case studies concerned with Estonian in Germany, in those concerning the minority language groups in Russian Karelia and Norway as well as in the case study on Meänkieli speakers in Sweden, the obtained data show a clear bias in favour of language activists as respondents.

In an ideal case, all study populations should have been divided equally by gender and four rough age categories (18–29, 30–49, 50–64, and 65+) with as much variation as possible in terms of the demographic background of the respondents. In those case studies in which the sampling frame was an official register, it was possible to obtain the intended division by gender and age, while in the majority of case studies this ideal had to be eased, as no official registers were available for sampling. According to the initial interview design, the interviewees in the age cohort focus groups should have been selected mainly from those who had returned the survey questionnaire; the idea was to ask in the questionnaire whether the respondent will allow a further contact and is willing to give an interview. In practice, however, this recommendation could be followed in only three case studies (Seto, Võro, Meänkieli), and even in these only partially: in all case studies at least a part of the interviewees had not participated in the survey. One of the reasons for these deviations from the pre-planned procedure was that although there were survey respondents who had announced their willingness to be contacted for an interview, they mostly belonged to the two oldest age groups. Furthermore, in most cases the participants of the focus group interviews were selected by the responsible research teams, usually with the help of the minority organisations, and the selection was based on researchers' scholarly knowledge of the local circumstances and networks within the minority and the majority communities. In some case studies, Facebook and Myspace were used to complement the selection of interviewees.

When compiling the minority focus groups, participants with at least a receptive ("passive") command of the minority language were preferred. The control group interviewees were selected from people engaged in decision-making bodies, an additional criterion was that they had shown some interest – positive or negative – to matters concerning the investigated minorities.

3.3. Case-specific deviations in the modes of data collection

According to the original project plan, the new data were to be gathered in late 2010 to early 2011 using mail surveys addressed to randomly selected respondents, and by semi-structured focus group and individual interviews which should have taken place during the spring of 2011. In practice, however, the data collection modes and the timeframes were only followed as planned in the case studies involving Hungarian in Austria and Slovenia and the two case studies in Finland, namely, Estonian and Karelian.

In Estonia and Finland, the data were collected by mail surveys between January and March 2011. In Estonia, however, the questionnaire surveys among Võro and Seto speakers were not carried out as a mail survey but the Estonian research team thought it best to change the survey mode to interviewer supported self-completion; the control group survey, however, was conducted as a mail survey.

In Norway, the mail survey addressed to the Kvens was launched as planned in October 2010, but due to problems in acquiring a proper sampling frame, the survey could only start after a maximally representative sample had been obtained from the Norwegian population register. The survey was finally conducted between April and June 2011; however, due to the extremely low response rate, an additional web survey was conducted in summer 2011. The web survey also yielded a very low response rate and for this reason the web responses are not included in the Kven data set.

In Sweden, the mail surveys among Meänkieli speakers and Finnish speakers were delayed due to organisational problems at the University of Stockholm. The empirical data were collected there between February and May 2011.

The most divergent data collection modes were adopted in Russia and in Germany. In the Republic of Karelia, the data collection ended up being a fairly heterogeneous combination of different sampling procedures. Speakers of Veps and Karelian were first screened by local researchers from larger groups of people; in practice this meant that the research teams made sure that those who were selected as participants in the survey as well as those to be interviewed, actually were

capable of using Karelian or Veps. As screening was not characteristic of any other case study, the Karelian and the Veps data are not strictly comparable with the other data sets. After the respondent screening, the actual data collection was conducted using interviewer-aided self-completion, that is, fieldworkers helped the respondents fill in the questionnaires; for example, by providing explanations of what was meant by the questions and by translating parts of the questionnaire when needed. In Russia, the data collection took place in March 2011.

In Germany, the search for potential Estonian speakers was initiated already in spring 2010, but recruiting participants proved to be very difficult. One reason for this was that as Estonians in Germany constitute a very small group, guaranteeing the anonymity of the respondents posed a considerable challenge. As a result, many individuals did not want to participate as they feared being identified by others later. Ultimately, the respondents in the case study of Estonian in Germany consisted of members of Estonian associations who volunteered to participate and of members of a selection of Facebook groups who reacted to an open call inviting people to participate in a mail survey. The data collection among Estonians in Germany was conducted between November 2010 and May 2011.

4. Digital formats of the data sets, data processing, and editing

All the data collected by the surveys and tape- and video-recorded in the interviews were processed into a digital format. The interviews were transcribed employing a rough transcription system and are included in the database as Transcriber files. For project-internal purposes, the interview data also were analysed using ELAN; the encoded ELAN files were not included in the database but remained for personal use by members of the case-specific research teams.

The survey data sets include only the results of the closed questions, since only these could provide numerical values automatically. The entire ELDIAdata minority-language database contains in total 3 388 individual records; the language-specific data sets include 340 variables each. The entire control group database contains in total

1 460 records from seven countries; each country-specific data set covers 280 variables.

All values in all data sets have been checked by statisticians. In order to make it possible to scan the responses directly into individual data sets, the survey questionnaires were typeset in a coordinated manner. The scanning process was decentralised among various participating organisations while the optical character recognition (OCR) of data sets took place at the University of Vienna. In the databank the survey-questionnaire data are stored in a format which allows for computer-based processing with a wide variety of statistical applications.

The basic data sets and analysis were programmed using the SAS software. Later on, the data sets were transformed into the SPSS data format as well. Although the coordination of the digitalisation process of the survey data by the ELDIA teams in Mainz and Vienna was mostly successful, some problems occurred due to technical difficulties with the optical character recognition (OCR) software and due to inconsistencies with the questionnaire contents. Hence, after the questionnaires were scanned (or entered into the required form manually), a substantial effort had to be taken to edit all data sets as uniformly as possible for comparative analyses. The editing was conducted by the ELDIA statistics team in Helsinki. The first edition round was completed in September–October 2011, the second round which provided the final data files now included in the ELDIAdata was conducted in June–July 2013.

The data editing process revealed a few technical errors in the comparability of the questionnaires and hence in the comparability of the survey data across all data sets. Some of the corrections have to be taken into account when using certain parts of the Kven and the Meänkieli data sets and the control group data sets from the Russian Federation and Estonia. These data sets are provided with explanations and instructions for future users.

5. Availability of the ELDIA data and the conditions for using these in further research

The ELDIAdata Database is managed by the University of Mainz and a Board of Administration consisting of representatives of the research institutions that participated in ELDIA. The database is available for research purposes only and only per a written request addressed to the Board. The EuLaViBar Toolkit (Spiliopoulou Åkermark et al. 2013), containing a revised version of the ELDIA questionnaire, templates and explanations of the statistical method, can be downloaded from the PHAIDRA repository (https://phaidra.univie.ac.at/o:301101).

References

Laakso, Johanna, Anneli Sarhimaa, Sia Spiliopoulou Åkermark & Reetta Toivanen 2013: *Summary of the Research Project ELDIA (European Language Diversity for All). Abridged version of the original English-language report*. Available at: ‹https://fedora.phaidra.univie.ac.at/fedora/get/o:304813/bdef:Content/get›

—— 2016: *Towards openly multilingual policies and practices. Assessing minority language maintenance across Europe*. Linguistic Diversity and Language Rights 11. Bristol: Multilingual Matters.

Spiliopoulou Åkermark, Sia, Johanna Laakso, Anneli Sarhimaa, Reetta Toivanen, Eva Kühhirt & Kari Djerf 2013: EuLaViBar Toolkit. Wien: ELDIA. Available at: ‹https://phaidra.univie.ac.at/o:301101›