http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1473

Web-based Exploration of Results From a Large European Survey on Dictionary Use and Culture: ESDexplorer

Sascha Wolfer, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (wolfer@ids-mannheim.de)*

Iztok Kosem, *Department of Translation, Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia* (*iztok.kosem@ff.uni-lj.si*)

Robert Lew, *Faculty of English, Department of Lexicography and Lexicology, Adam Mickiewicz University, Poznań, Poland (rlew@amu.edu.pl*)

Carolin Müller-Spitzer, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (mueller-spitzer@ids-mannheim.de)*

Maria Ribeiro Silveira, *Institute for the German Language (Institut für Deutsche Sprache), Mannheim, Germany (ribeiro@swhk.ids-mannheim.de)*

Abstract: We present ESDexplorer (https://owid.shinyapps.io/ESDexplorer), a browser application which allows the user to explore the data from a large European survey on dictionary use and culture. We built ESDexplorer with several target groups in mind: our cooperation partners, other researchers, and a more general public interested in the results. Also, we present in detail the architecture and technological realisation of the application and discuss some legal aspects of data protection that motivated some architectural choices.

Keywords: SURVEY, DATA COLLECTION, DATA PROCESSING, DATA PRESENTATION, DATA ANALYSIS, TECHNOLOGY AND ARCHITECTURE, TARGET GROUP, PLOT, BROWSER APPLICATION, ESDEXPLORER

Opsomming: Webgebaseerde verkenning van die resultate van 'n omvattende Europese opname van woordeboekgebruik en -kultuur: ESDexplorer. Ons stel ESDexplorer (https://owid.shinyapps.io/ESDexplorer), 'n webblaaiertoepassing wat die gebruiker toelaat om die data van 'n omvattende Europese opname van woordeboekgebruik en -kultuur te verken, bekend. Met die bou van ESDexplorer het ons verskeie teikengroepe in gedagte gehad: ons samewerkingsvennote, ander navorsers, en 'n meer algemene publiek wat in die resultate sou belangstel. Ons bespreek ook die argitektuur en tegnologiese totstandkoming van die toepassing in

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1473

besonderhede en brei uit oor enkele regsaspekte rakende databeskerming wat sommige argitektuurkeuses gemotiveer het.

Sleutelwoorde: OPNAME, DATAVERSAMELING, DATAVERWERKING, DATA-AANBIEDING, DATA-ANALISE, TEGNOLOGIE EN ARGITEKTUUR, TEIKENGROEP, GRAFIEK, WEBBLAAIERTOEPASSING, ESDEXPLORER

1. Introduction

On 8 May 2017, a large-scale survey on dictionary use[1] was launched in 26 European countries and Brazil[2]. The main goal of the survey was to provide an up-to-date picture on dictionary use and culture (particularly) across Europe. This has been by far the most extensive dictionary-related survey to date, both in terms of the sheer number of participants and in terms of the breadth of coverage along national and linguistic dimensions. Due to its large scale, the survey presented particular challenges with regard to data collection, processing, and presentation. A core group of four researchers (Iztok Kosem, Robert Lew, Carolin Müller-Spitzer and Sascha Wolfer) drafted the general part of the survey. The general part consisted of 13 questions that were accompanied by 11 questions eliciting personal data from the participants (henceforth referred to as "meta-variables"). Around 60 researchers all over Europe (so-called "local partners") translated this original English version of the questionnaire into their local languages and helped to disseminate the survey in their countries. After all the translations were completed, different language versions were implemented in the online survey system Unipark Questback at the Institute for the German Language in Mannheim.

The local partners were given the opportunity to create local parts of the survey in their native language consisting of up to five short questions. These local parts were only presented to the participants from the respective countries and are not covered in this contribution or available in ESDexplorer.

Between 8 May and 9 July 2017, 9,373 participants completed the survey. Figure 1 shows the distribution of participants over countries and the professional status of the participants. All data is accessible in raw format to the core group. The local partners were given access to the raw data[3] from participants from the respective country and — if present — the raw data from their local part. An analysis of the survey data (Kosem, Lew, Müller-Spitzer, Ribeiro Silveira, Wolfer et al. 2018) and one article in German, mainly covering the German local part, has already been published (Müller-Spitzer, Ribeiro Silveira, Wolfer, Kosem and Lew 2018). A Slovene paper by Arhar Holdt (2018) focuses on the Slovene perspective.

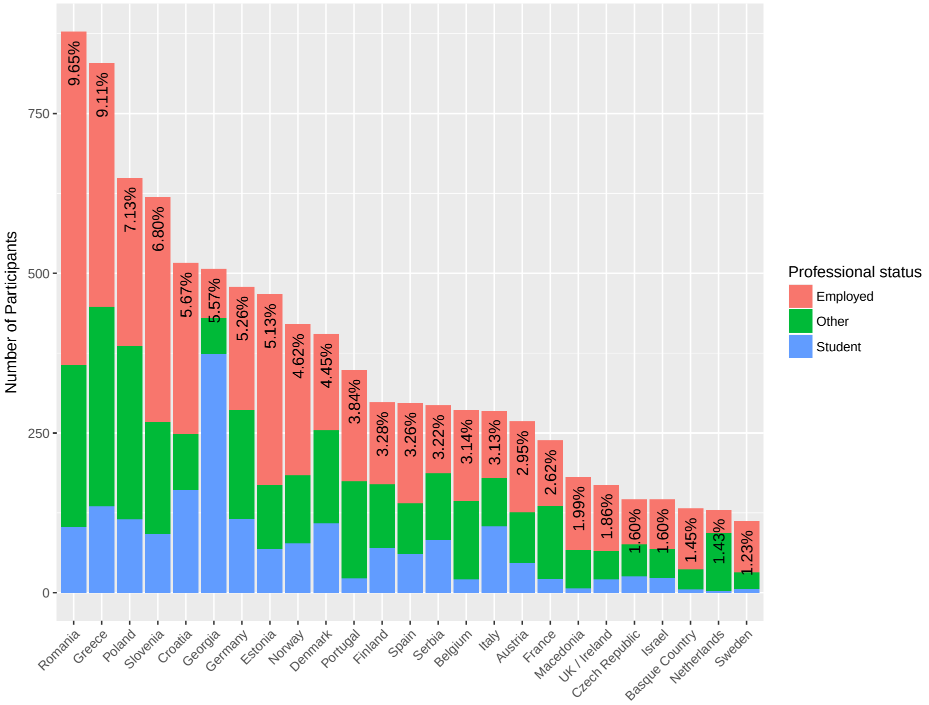


Figure 1: Number (*y*-axis) and percentages of participants per country. The bars are divided by professional status of the participants ("Student" does not contain "Ph.D. student", Ph.D. students are counted as "Other").

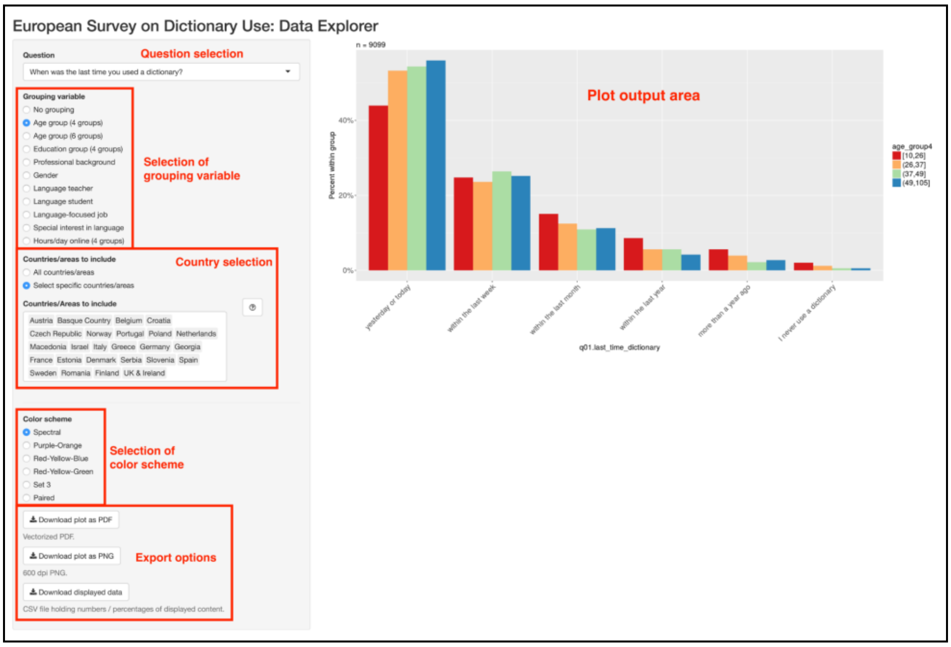
In the following section, we introduce ESDexplorer from the perspective of the user. In section 3, we go into more detail regarding the groups that we had in mind when designing ESDexplorer. Section 4 describes the technology that was used in building ESDexplorer, and briefly explains the technical mechanism of the server-side calculations which are not visible to the user. In section 5, we conclude this contribution with a summary.

2. ESDexplorer

The application is available under https://owid.shinyapps.io/ESDexplorer. In ESDexplorer, data from 11 questions from the general part of the survey is accessible in aggregated form. Due to the declaration of consent given by the participants, we are not permitted to make available the raw data which, in principle, could be used to trace back answers to individuals[4].

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1473

Please refer to Figure 2 for an overview of ESDexplorer's user interface, with labels identifying its main elements.



European Survey on Dictionary Use: Data Explorer

Question: When was the last time you used a dictionary?

Grouping variable
○ No grouping
● Age group (4 groups)
○ Age group (6 groups)
○ Education group (4 groups)
○ Professional background
○ Gender
○ Language teacher
○ Language student
○ Language-focused job
○ Special interest in language
○ Hours/day online (4 groups)

Countries/areas to include
○ All countries/areas
● Select specific countries/areas

Countries/Areas to include: Austria, Basque Country, Belgium, Croatia, Czech Republic, Norway, Portugal, Poland, Netherlands, Macedonia, Israel, Italy, Greece, Germany, Georgia, France, Estonia, Denmark, Serbia, Slovenia, Spain, Sweden, Romania, Finland, UK & Ireland

Color scheme
● Spectral
○ Purple-Orange
○ Red-Yellow-Blue
○ Red-Yellow-Green
○ Set 3
○ Paired

Vectorized PDF.

600 dpi PNG.

CSV file holding numbers / percentages of displayed content.

Figure 2: An annotated screenshot of ESDexplorer's user interface. The annotated labels appear in *italics* in the text.

The left-hand side with the greyish background is reserved for user input, whereas the right-hand side of the screen (*Plot output area*) presents the output given by the system. The input area on the left-hand side includes a number of elements. First, the user needs to select a specific question (*Question selection*). Eleven questions from the general part are available. Two questions from the general part are not available for selection. The first of these was an open-ended question asking participants which monolingual dictionaries they used. Such an open question has to be coded manually before the results can be meaningfully presented in a visual format; for example, responses such as "oed.com", "www.oed.com", "OED online", "oed on the web" and so on need to be mapped to a single entry[5]. The other question not currently available for analysis in ESDexplorer is "How much are you willing to spend on a good monolingual dictionary of [your language] (in [your currency])?". Here, currency conversion would have to be included to obtain reasonable results. Also, other factors like variation in purchasing power between the participating

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1473

countries may need to be considered. Since this has not yet been done, we decided to exclude this question from ESDexplorer.

Optionally, data analysis can be grouped by a meta-variable describing the status of participants (*Selection of grouping variable*). For the age of participants, two granularities (four and six groups) are available. Two meta-variables that were included in the survey are not available for analysis in ESDexplorer: native language and device usage (participants were asked to indicate all devices they used on a daily basis; the options given were: desktop computers, laptops, tablets, and smartphones. We did not include native language because the list of available native languages was very long (44 items). A visualization with that many categories would not be legible at all. Device usage was not included because there is no straightforward way to represent all the different combinations that are possible for the four options. Since one of the main goals of ESDexplorer is to represent the information in a clear and compact way, it seemed like a good decision not to include this meta-variable. If "No grouping" (first option) is selected, the bars in the output plot are collapsed to one grey bar per answer category (*x*-axis) and the *y*-axis switches to counts instead of percentages within the group. The overall percentages are then also annotated above the bars.

If the user wants to exclude certain countries from their analysis they can do so with the *Country selection.* In the selection list, any subset of countries can be selected. The "n = [number]" above the plot tells the user how many participants contributed their data to the current plot. Consequently, this number decreases when fewer countries are selected.

If a grouping variable is selected, the user might choose between six different color schemes (*Selection of color scheme*) to accommodate the context in which the plot might be used. Users can export the data currently shown in the *Plot output area* with three *Export options*. PDF export provides a vector-based graphic that can be scaled to any size. The second download option is a high-resolution (600 DPI) PNG file. With the last option, the user can download a comma-separated data file containing all the counts or percentages that underlie the plot currently shown.

3. Motivation and target groups

We had three groups of users in mind when designing ESDexplorer. The initial idea to create the application was to help those local partners that do not have training in data representation, manipulation and analysis to access the data in an easy and straightforward way. The application could thus serve as a starting point for our local partners to conduct preliminary analyses that might lead to publications in their local language. With ESDexplorer, the partners can access the data from the general part of the survey and use the plots generated by the application to document and compare answers from any combination of countries. The plots generated by the system can be used for their own publications,

http://lexikos.journals.ac.za; https://doi.org/10.5788/28-1-1473

either directly (as PNG files) or with little extra manipulation, e.g. using the CSV files with Excel or similar software.

The second group that we had in mind is a broader scientific community. Researchers from outside the consortium that co-operated in the survey might check the plausibility of more fine-grained analyses presented in publications that are based on this data. It has to be said, though, that this does not satisfy the broadest requirements for reproducible research to the full extent. To do so, we would have to provide all the data on an individual level, i.e. what we referred to as "raw data" above. However, due to data protection issues and the declaration of consent we asked our participants to acknowledge, we are not allowed to make available the raw data to anyone who was not part of the consortium. Nevertheless, we believe that ESDexplorer at least allows for detailed plausibility checks of analyses that are presented elsewhere (e.g. in journal papers).

The third target group is the broader, non-scientific audience that might be interested in lexicographic research. With ESDexplorer, these users have an easy-to-use point-and-click interface where they can learn something about the culture and use of monolingual dictionaries in their own country or all over Europe. It may also be the case that the participants of the study are interested in the final results. With the application, they can explore the data by themselves without the researchers functioning as "gatekeepers".

Our online visualization system can also have a didactic application beyond lexicography: ESDexplorer might serve as a model example for university teachers to illustrate the visualization possibilities for questionnaire studies. Since the application shows the results of a questionnaire study, it might nicely complement theoretical discussions about questionnaire design.

4. Technology and architecture

ESDexplorer is built using the R (R Core Team 2018) package "shiny" (Chang et al. 2017). With this package, one can build web applications without extensive web development skills. The basic architecture of a Shiny application consists of two scripts written in R code. One script controls the behaviour of the user interface with the input elements (so-called "widgets") and outputs (mostly plots and downloadable data). All the input widgets that are used in ESDexplorer come with a standard Shiny installation and can be readily used. The other script determines what is going on "behind the scenes". Here, the developer controls data management and statistical computation on the server. The computation underlying the graphs is not done in the user's browser but almost exclusively on the server itself. The only computations that are running in the user's browser are for showing the output and getting user input from the widgets and passing them along to the server.

Interestingly, the server script uses a data set that is very close to the raw data, i.e. the data on an individual level. This data set is used to aggregate the data so that it can be displayed in the graph that the user requested. Due to the encapsulated nature of the computations and the raw data itself, the user cannot access this raw data file (nor can they access the server script, but this is less critical). This is necessary because, as indicated in section 1, we are not allowed to disclose the individual data.

Whenever a parameter is changed on the left-hand side (i.e. the user input section), the plot is updated. This is thanks to the reactive nature of the Shiny environment: whenever the user changes something in the user interface, the server script detects this change and a new calculation is triggered. For very large data sets (e.g., with several million cases or a large number of variables), this process might be slow. But with our dataset of 9,373 rows (= participants) and roughly 100 columns (= variables), this is no problem for real-time server-side calculations.

Luckily, the Shiny environment comes with its own session management system, so the developer of the application does not have to deal with the challenge of several users accessing the application at once.

While Shiny and R itself are free software, a Shiny application still has to be hosted on a Shiny server, so that users can access it online. At the moment, ESDexplorer is hosted with shinyapps.io, a service provided by RStudio, the company that also created the Shiny package. This is a proprietary service with a free tier. This free tier, however, only includes very limited usage. Hence, we chose the cheapest paid option to host ESDexplorer at the moment. An alternative is to host your own Shiny server using the open-source Shiny server, which is also available from RStudio. ESDexplorer might be transferred to such a solution in the long term.

5. Summary

ESDexplorer is a browser application where users can explore the results of the 2017 European Survey on Dictionary Use. The users can use grouping variables in their analysis and subset the data by country. With the application, we hope to reach three target groups: our local partners, the broader scientific community in lexicography and related disciplines, and the general public. ESDexplorer is implemented in Shiny, a framework for the dynamic and user-adaptive presentation of data.

Endnotes

1. A full list of participating researchers and countries can be accessed at http://www.elexicography.eu/events/european-survey-on-dictionary-use/ [last access on August 8th, 2018].
2. Brazil has been included primarily to be able to compare the Portuguese and Brazilian answers.

3. With "raw data", we refer to the individual questionnaires. Technically, the raw data is one large table with all completed questionnaires stored in rows and all the variables in columns.
4. Tracing back a specific questionnaire to an individual is still highly unlikely using the raw data. However, through a combination of country, native language, age, years of formal education and profession, it is theoretically possible. With aggregated data, it is definitely impossible to "track" single individuals.
5. Users gave 6,697 different answers (types) to this question (each user was allowed to enter five dictionaries). Altogether, 15,663 answers (tokens) were given. The OED example in the text would contain 4 tokens and 4 types that would have to be mapped to a single type by manual coding.

References

Arhar Holdt, Š. 2018. The Attitude of Language Users Towards General Monolingual Dictionaries: The Slovene Perspective. *Slovenščina 2.0,* 6(1): 1-36. Retrieved from http://slovenscina2.0.trojina.si/arhiv/2018/1/Slo2.0_2018_1_01.pdf.

Chang, W., J. Cheng, J. Allaire, Y. Xie and J. McPherson. 2017. shiny: Web Application Framework for R (Version 1.0.5). Retrieved from https://CRAN.R-project.org/package=shiny.

Kosem, I., R. Lew, C. Müller-Spitzer, M. Ribeiro Silveira and S. Wolfer et al. 2018. The Image of the Monolingual Dictionary Across Europe: Results of the European Survey of Dictionary Use and Culture. https://doi.org/10.1093/ijl/ecy022.

Müller-Spitzer, C., M. Ribeiro Silveira, S. Wolfer, I. Kosem and R. Lew. 2018. Eine europaweite Umfrage zu Wörterbuchbenutzung und -kultur: Ergebnisse der deutschen Teilnehmenden. *Sprachreport* 2 (2018): 26-35. Retrieved from http://pub.ids-mannheim.de/laufend/sprachreport/pdf/sr18-2.pdf#page=28.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.