



EUROPEAN LANGUAGE EQUALITY

D2.12

Report from Wikipedia

Author	Maria Heuschkel
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D2.12
Deliverable title	Report from Wikipedia
Type	Report
Number of pages	52
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP2: European Language Equality – The Future Situation in 2030
Task	Task 2.2 The perspective of European LT users and consumers
Author	Maria Heuschkel
Reviewers	Natalia Resende, Tea Vojtěchová
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	<p>European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland</p> <p>Prof. Dr. Andy Way – andy.way@adaptcentre.ie</p> <p>European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany</p> <p>Prof. Dr. Georg Rehm – georg.rehm@dfki.de</p> <p>http://www.european-language-equality.eu</p> <p>© 2022 ELE Consortium</p>

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language “Prof. Lyubomir Andreychin”	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ülikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

1. Introduction	1
1.1. About Wikimedia Deutschland and the Wikimedia movement	2
2. Methodology and Instruments	3
2.1. Online Survey	3
2.2. Interviews	4
3. Analysis of Survey Responses	5
3.1. Survey responses	5
3.1.1. Respondents' profiling	5
3.1.2. Language Coverage	7
3.1.3. Evaluation of the Current Situation	8
3.1.4. Predictions and Visions for the Future	11
3.2. Interview Responses	13
3.2.1. Gaps and Problems	13
3.2.2. Lack of language experts in the Wikimedia Community	14
3.2.3. Recommendations	15
3.2.4. Wikidata, Lexemes and Abstract Wikipedia	16
3.2.5. Initiatives for building audio corpus for languages/ Oral Knowledge . .	17
4. Conclusions	18
A. LT users and consumers survey	21
B. Tables for Analysis	38
C. Additional tables and graphs	42

List of Figures

1. In which country are you based?	6
2. Which of the following best describes the type of organisation you work for? .	7
3. Which of the official European language(s) listed below do you or your organisation work with?	8
4. Which language technology tools/applications listed below do you or your organisation use with the official European language(s) you or your organisation work with?	9
5. Please choose the option that best describes the level of language technology support for the language(s) your community or your organisation work with. .	12
6. Please indicate the best option that describes your vision for the future of languages technology.	13
7. Full survey as published (page 1/18)	21
8. Full survey as published (page 2/18)	22
9. Full survey as published (page 3/18)	23
10. Full survey as published (page 4/18)	24
11. Full survey as published (page 5/18)	25
12. Full survey as published (page 6/18)	26
13. Full survey as published (page 7/18)	27
14. Full survey as published (page 8/18)	27
15. Full survey as published (page 9/18)	28
16. Full survey as published (page 10/18)	29
17. Full survey as published (page 11/18)	30
18. Full survey as published (page 12/18)	31
19. Full survey as published (page 13/18)	32
20. Full survey as published (page 14/18)	33
21. Full survey as published (page 15/18)	34
22. Full survey as published (page 16/18)	35
23. Full survey as published (page 17/18)	36
24. Full survey as published (page 18/18)	37

List of Tables

1. Type of survey questions	3
2. Wikipedia language versions for the ELE languages, retrieved from https://wikistats.wmcloud.org/display.php?t=wpNovember2021	39
3. Number of recordings on Lingua Libre for languages of the ELE consortium, retrieved from https://lingualibre.org/wiki/LinguaLibre:Stats/Languages on November 9 2021	40
4. Number of Lexemes in Wikidata for the languages of the ELE consortium, retrieved from https://ordia.toolforge.org/language on December 1 2021	41
5. List of Wikimedia chapters and user groups in the EU and dealing with European languages	42
6. Breakdown of answers count to the question “In which country are you based?”	42
7. Breakdown of answers count to the question “Which of the following best describes the type of organisation you work for?” (Example of mandatory single choice question)	42
8. Breakdown of answers to the question “Which community are you representing? (e. g., Wikidata, Italian Wikipedia, User Groups etc)?”	43

9. Breakdown of answers to the question “For which language (s) you, your community or your organisation use language technology tools (e. g., Translation tools, Spell/grammar checkers, web search engines, social media, language learning tools)?”	44
10. Breakdown of answers count to the question “Which tools/applications do you use with these languages?”	45
11. Breakdown of answers to the question: “Which tools/applications do you use with these minority/regional/lesser-used languages? if “other”, please specify.”	45
12. Full list of answers to “Which tools or applications that substantially use language technology do you want to see in the community you represent that are not available today? (we welcome any suggestion, even ideas that are not possible with current technology)?”	46

List of Acronyms

EC	European Commission
EU	European Union
ELE	European Language Equality (<i>this project</i>)
ELE Programme	European Language Equality Programme (<i>the long-term, large-scale funding programme specified by the ELE project</i>)
CEE	Central and Eastern Europe
GLAM	Galleries, Libraries, Archives
Museums	
LOD	Linked Open Data
LR	Language Resource/Resources
LT	Language Technology/Technologies
SME	Small and Medium Sized Enterprise
SPARQL	SPARQL Protocol and RDF Query Language
SRIA	Strategic Research Innovation and Employment Agenda
WikidataCon	Wikidata Conference
WMDE	Wikimedia Deutschland

Abstract

The following report is part of the consortium of the European Language Equality (ELE) Project funded by the European Commission (EC). The results presented have been acquired through a survey, discussion rounds, interviews as well as literature review. The goal of the document is to report on the view of the Wikimedia communities on the state of Language Technology (LT) in Europe. The report has identified several gaps, problems and needs related to working with smaller, minority, regional or under-resourced European languages from the perspective of Wikimedia communities. This input will be used to create a strategic roadmap and agenda to achieve Digital Language Equality in Europe by 2030.

1. Introduction

This document reports on the findings of a consultation with representatives from the Language Technology (LT) users and consumers within the Wikimedia community, conducted by the EU project European Language Equality (ELE). These results will serve as input for a strategic research, innovation and deployment agenda (SRIA) and roadmap, in order to tackle the striking imbalance between European languages in terms of the support they receive through LTs by 2030.

The ELE project sought to collect the views of European LT users and consumers and to consolidate their perspective on the differences in terms of technologies for the languages they work with and of the measures that need to be put in place so that all European languages are equally supported through technology by 2030.

Due to the interdisciplinary nature of the field of LT, which stands at the intersection of Linguistics, Computational Linguistics, Computer Science and Artificial Intelligence, the ELE project brings together diverse groups of stakeholders including researchers, representatives of communities of LT users and consumers, language professionals (e. g., translators, lecturers and professors in the field of Linguistics and Computational Linguistics) and stakeholders from different economic sectors (e. g., banking, health).

Although the methodology and instruments used in gathering feedback from LT users have been common to other ELE consortium members who were responsible for spreading the survey in their communities, this report covers and analyses the subset of responses of stakeholders contacted by Wikimedia Deutschland.

The Wikimedia projects, and more specifically, the different language version Wikipedias are considered an important tool for language revitalization and preservation projects. Moreover, these various Wikipedias have provided much needed training data for many of today's data-driven Natural Language Processing (NLP) tools and language models. In terms of multilingual digital content, linked data and an organically growing resource, Wikipedias (or any other Wiki projects) are invaluable components of the future of language technology.

Wikimedia communities around the world are dedicated to making content available in every language, especially indigenous, small and under-resourced languages – free and open for everyone. However, creating and maintaining a Wikipedia can be hard and especially so for smaller communities with only a small number of speakers. Maintaining their own language version Wikipedia can be challenging and time intensive (Norge, 2021). Nonetheless, the Wikimedia movement wants to make a difference for small, minority, regional, lesser used and under-resourced languages and provide opportunities for those language communities to contribute to and work with their languages in an online environment. As such, the participation of Wikimedia Deutschland as one of 52 partners in the ELE project (2021/22) allows the voices of contributors to the Wikimedia projects to be heard.

The following report will include an analysis carried out on a survey that was shared across

the Wikimedia communities, along with interviews with stakeholders, as well as past discussion groups on the topic of language diversity. The main focus of the data collection was the perspective of the Wikimedia communities on Language Technologies and Digital Language Equality in Europe. Following an introduction to Wikimedia, the methods of the report (both quantitative and qualitative) are explained. In the course of the analysis the survey answers, interviews with stakeholders and literature review are included.

1.1. About Wikimedia Deutschland and the Wikimedia movement

Wikimedia is a global movement to promote free knowledge. Like Wikipedia, this movement grew through volunteer efforts to make the sum of all knowledge freely accessible. Wikipedia, the free encyclopedia, is the first and most successful of many projects within the Wikimedia family. Every day, tens of thousands of volunteers around the world are working to improve Wikimedia projects. All of these projects are operated by the non-profit Wikimedia Foundation in San Francisco. Worldwide, 40 independent chapters support Wikimedia at the national level. The non-profit organisation Wikimedia Deutschland is the oldest and largest among them. The association was founded in 2004 to promote free knowledge. Our goal is to support Wikipedia and its sister projects, and to promote the concept of free knowledge by following their example: the primary objective is to provide free access to and free reuse of the sum of all knowledge. That is our understanding of the basic human right to education. The association's work in pursuit of these objectives is funded by donations. The Wikimedia movement is the global community of people that want to contribute to free and open knowledge through the Wikimedia projects. Wikimedia in Europe and the EU consists of both organized and official chapters, as well as a more flexible organisation form of user groups and many more community groups or community members that have no official affiliates status and nonetheless are an integral part of the Wikimedia movement.

The Wikimedia movement currently (November 2021) has 38 chapters, 136 user groups and 2 thematic organisations (WikimediaMovement, 2021), see Table 5.

The heart of the Wikimedia movement is the communities working on the Wikimedia projects. The most well-known of those Wikimedia projects is Wikipedia, the Online Encyclopedia. Further open access projects that are part of the Wikimedia movement are Wiktionary (dictionary & thesaurus), Wikiquote (quotations), Wikibooks (textbooks & manuals), Wikinews (open journalism), Wikisource (source texts), Wikivoyage (travel guide) and Wikiversity (learning resources), Wikimedia Commons (media files), Wikidata (knowledge base), MediaWiki (Wiki software development) and Wikispecies (species directory). All those projects are open content projects, available in a different number of languages or including multilingual content (Wikidata and Wikimedia Commons are multilingual projects).

There are currently 314 active Wikipedia language editions (Movement, 2021a). Table 2 shows an overview of the language versions Wikipedias covered by the ELE consortium. The five biggest language versions (measured by the number of articles) are the English, Cebuano, Swedish, German speaking and French speaking Wikipedia. As for the Swedish Wikipedia having such a high article count, the explanation for this can be found in the automated bot that has produced so-called stubs (short Wikipedia articles) in the Swedish Wikipedia¹. Interestingly, there is also a rather large number of smaller and minority languages that are being represented in their own language version (e. g., Cornish, Manx, Sorbian, Saami languages).

As for the Wikimedia projects other than Wikipedia, we also see the case of the “big languages” having a lot of content, first and foremost the English and French Wiktionary (Movement, 2021b). The next cluster of languages and projects that are being covered by the ELE project (with a total page count between 10,000 and 8,000) are the German Wiktionary, the

¹ <https://de.wikipedia.org/wiki/Lsjbot>

Polish Wikisource, the Serbo-Croatian Wiktionary and Spanish Wiktionary as well as the English Wikisource and Swedish Wiktionary.²

The interest to work with languages and keep languages alive is deeply embedded in the Wikimedia movement. There are numerous initiatives within the movement that explicitly deal with languages, e. g., the annual Wikimedia language conference Celtic Knot.³

2. Methodology and Instruments

2.1. Online Survey

The survey was addressed to LT users and consumers, and sought to elicit the respondents' views in a way that would facilitate the analysis, consolidation and integration of the collected feedback into the ELE SRIA and roadmap. It had 63 questions in total. Some of the questions depend on previous answers. As a result, a respondent was presented with a minimum of 30, to a maximum of 63 questions, including the catch-all "if other" questions. 46 questions were mandatory, of which 33 were closed questions (single or multiple choice). Table 1 shows an overview of the types of questions.

Question types	Mandatory	Optional	Total
Closed	20	13	33
Open-ended	26	4	30
Total	46	17	63

Table 1: Type of survey questions

The survey was structured in four main parts. If any of the provided answers were not applicable, the respondents had the option to enter a different answer through the option "if other, please specify".

- **Part A. Respondents' profiling:** the first part of the survey included 13 questions for the demographic profiling of respondents with an emphasis on characteristics relevant to the task at hand, i. e.,
 - Country in which respondents are based
 - Name of the organisation/representative body for which respondents work
 - Communities they represent (if applicable)
 - Type of organisation for which respondents work
 - Sectors or domains in which respondents are active (if applicable)
 - Role of respondents in the organisation (if applicable)
 - Organisation's estimated revenue (if applicable)
- **Part B. Language coverage:** This part looked into the European languages the respondents work with and the languages they intend to include in their workflow, i. e.,
 - Languages used by the organisations, associations, communities, professionals of LT users

² If there is a community that wants to launch a new language version Wikimedia project (Wikipedia or other), the final decision is made by the Wikimedia Foundation language committee: https://meta.wikimedia.org/wiki/Language_committee. This process is public and relies on input from the overall Wikimedia community.

³ https://meta.wikimedia.org/wiki/Category:Celtic_Knot_Conference

- Languages intended to be supported in the short- or medium-term
- **Part C. Evaluation of current situation:** Respondents were requested to evaluate the level of technology support for the official European languages they work with and that of any minority, regional or lesser used language, i. e.,
 - Differences in availability of LTs between the official European languages they work with and, if applicable, differences in availability of LTs between the minority, regional or lesser-used languages they work with
 - Gaps perceived in the technologies, tools or applications that respondents work with in relation to specific languages
 - Respondents' opinion in relation to performance of LTs with regard to specific languages
- **Part D. Predictions and visions for the future:** respondents are requested to share their needs and wishes for the future of language technologies, i. e.,
 - Policies or instruments that could contribute to speed up the effective deployment of LT in Europe equally for all languages
 - Prediction of future opportunities for LT in basic and applied research (scientific vision), in innovation and in industry
 - Expectations of the community with regards to the challenges an ELE Programme can address by 2030

Follow-up: The last three questions requested the respondents permission to be contacted for an interview and, given an affirmative answer, their contact details. Respondents were also requested to click on a confirmation question stating “By clicking on ‘Submit’, I agree that my personal data (email address and/or name) can be used according to the Privacy Policy of the European Language Equality (ELE) project”. The survey was designed, set up and published on the EU Survey platform.⁴ The full survey, as published online, is presented in Appendix A (p. 21 ff.). The survey link was distributed by Wikimedia through several emails to members of the Language community within Wikimedia movement by using several mailing lists.⁵ We also personally contacted 150 individual community members that are active in European language Wikipedias, Wikidata and Lexemes, Lingua Libre or other Wikimedia projects. It was announced at the Wikimedia Language conference the Arctic Knot in June 2021 (Heuschkel, 2021) and in the Wikimedia Newsletter in August 2021. It was additionally advertised through the European Language Equality and European Language Technology websites, LinkedIn page and Twitter account.

The survey was opened on 21 June 2021 and closed on 18 October 2021. In total, 246 responses have been collected, out of which 22 from respondents who were contacted by Wikipedia. This subset of responses, representing the views of the stakeholders contacted by Wikipedia is analysed in this report.

2.2. Interviews

In addition to the survey responses we have received from our community members, we carried out expert interviews with two Wikimedia members who had not filled in the survey. Both interviewees are long standing members of the Wikimedia movement. Interviewee 1 works with GLAM (Galleries, Librarians, Archives and Museums) to advance open access in institutions and Interviewee 2 works with minority languages and indigenous languages

⁴ <https://ec.europa.eu/eusurvey/runner/LTusers-consumers>

⁵ <https://lists.wikimedia.org/postorius/lists/languages.lists.wikimedia.org>

and in particular with the Northern Saami language and Skolt Saami. Those interviews took place in November 2021.

In addition, we organised discussion rounds at the WikidataCon 2021⁶ with a total of 25 participants.⁷ During the WikidataCon in October 2021, we asked the following questions:

1. What are we missing when working with small languages and under-resourced languages in Wikidata, Wikipedia and other Wikimedia projects? What do we need?
2. What do we want the world of lexicographical data and Wikidata to look like in 10 years for small, under-resourced languages? How do we get there?
3. What does “Digital Language Equality” mean to you (in terms of the language you speak and personally)?

During the conference for Central and Eastern Europe Communities in November 2021, we asked the following questions to participants:

1. What can the European Union and the European Commission do to support the digital survival of languages in the CEE communities?
2. What gaps and problems do you encounter when working with your languages online?
3. Which tools or applications that substantially use language technology do you want to see in the community you represent that are not available today?

In addition to the before mentioned interviews and discussion rounds, we also analysed internal documentation of discussions that had taken place in the past year on the topic of minority languages and language technologies in the Wikimedia movement.

3. Analysis of Survey Responses

In total, we managed to receive 22 responses from community members. From those 22 responses, many were not able to fill in all questions (especially those asking for the usage of specific LT in the languages or the performance of LT), the challenges with the research will be discussed in a separate section “Limitations and Challenges”.

3.1. Survey responses

3.1.1. Respondents’ profiling

We received answers from community members based in Spain (4), France (3), Germany (3), Russia (2), Hungary (2) as well as Slovak Republic (1), Ireland (1), Denmark (1), Bulgaria (1), Malta (1), Macedonia (1), United States (1) and Wales (1). The respondents were identified as community members of Wikipedia (14), Wikidata (5), Wiktionary (3), Lingua Libre (1) and Wikimedia Commons (1). See Table 6, Appendix C for full breakdown. Figure 1 illustrates the countries distribution.

The specific language versions for Wikipedia and Wiktionary that were mentioned in the survey were:

⁶ <https://pretalx.com/wdcon21/talk/NHRWTH/> as well as the Meetup of the Central and Eastern Europe Wikimedia community in November 2021.

⁷ https://meta.wikimedia.org/wiki/Wikimedia_CEE_Online_Meeting_2021/Programme/Submissions/Get_together!_Partnering_around_Languages_in_Europe_-_The_European_Language_Equality_Project

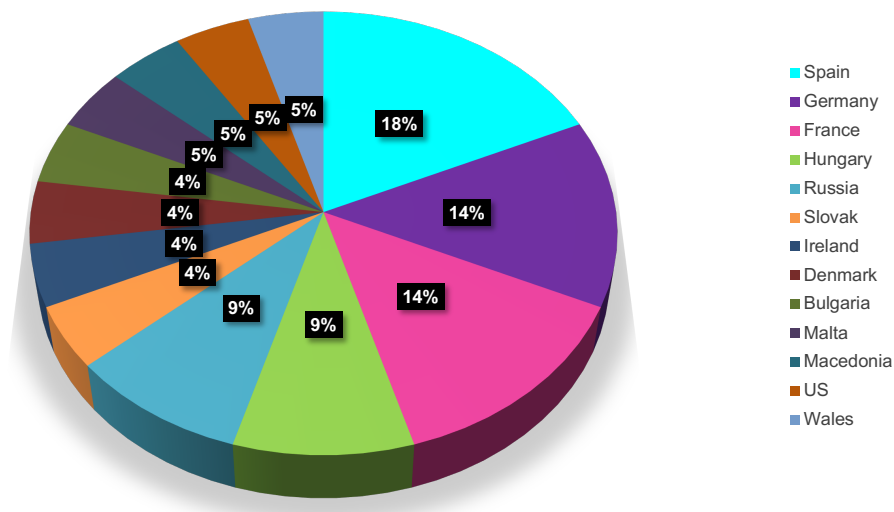


Figure 1: In which country are you based?

- Hungarian Wikipedia
- German speaking Wikipedia
- Slovak Wikipedia
- Basque Wikipedia
- French Wikipedia
- Catalan Wikipedia
- Macedonian Wikipedia
- Aragonese Wikipedia
- North Frisian Wikipedia
- Danish Wikipedia
- Bulgarian Wikipedia
- French Wiktionary
- Tacawit Wiktionary
- English Wiktionary

The respondents represented or are part of the following organisations:

- Wikimedians of Slovakia User Group
- Wikimedia Community of Saint Petersburg User Group

- Amical Wikimedia User Groups
- Wikimedia Foundation
- Universidad de Zaragoza
- CEE Spring User Group
- Technical University of Denmark
- Wikimedia Denmark
- Wales User Group

Most of the respondents have described the organization they are representing as Education/ research (11) with the second highest answer being N/A (6). This high number of N/A answers is not surprising as most of the community members are volunteers and the grey area between work for/ and volunteering was likely to make them choose N/A. The rest have stated that the type of organisation they work for is Other (2), Professional association (1), Large enterprise (1) and SME (1). See Table 7, Appendix C for full breakdown. Figure 2 shows the distribution of the types of organisations.

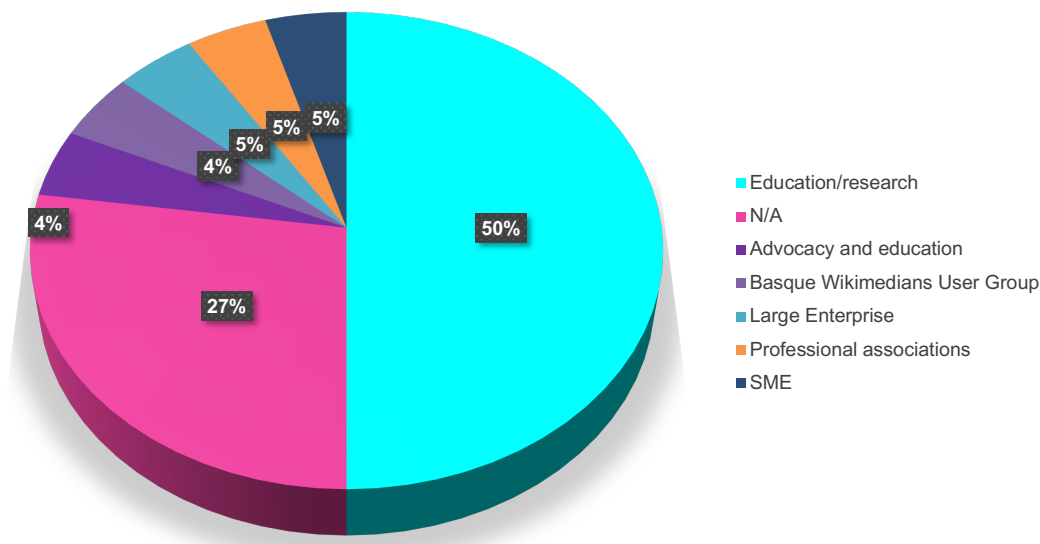


Figure 2: Which of the following best describes the type of organisation you work for?

8 respondents have stated they are volunteer contributors to the Wikimedia projects (e. g., editors, admins, patroller or bureaucrat). The rest of the respondents are secretaries, project managers (2), software engineers (2) and researchers/teachers (1). Most of the respondents feel like they are part of smaller organisations with 1-10 employees or volunteers (7) or 11-100 employees or volunteers (7). The rest of the respondents are distributed between organisations with 101-500 employees (3) and 501-5000 (3) or N/A (2).

3.1.2. Language Coverage

This section of the survey sets out to ascertain which languages are used at the respondents' workplace or organisation (either personally or globally within the organisation). It also

aims to capture if the scope of languages includes minority or regional languages and if it is set to broaden.

Figure 3 shows the distribution of the languages selected by respondents. For the languages with fewer mentions, please refer to Table 9 in Appendix C.

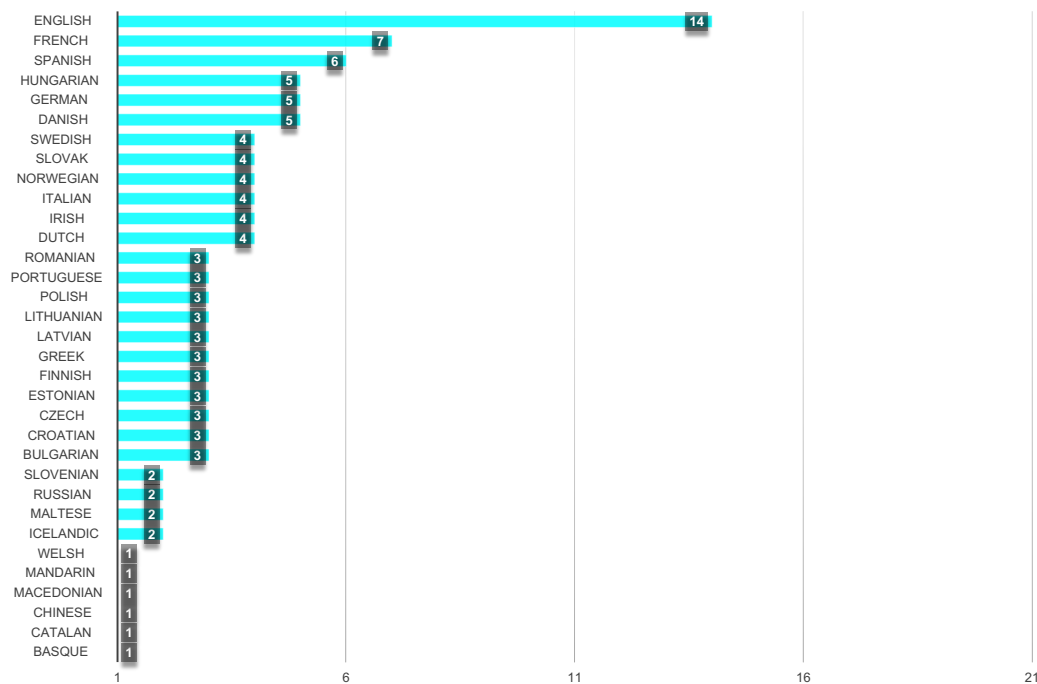


Figure 3: Which of the official European language(s) listed below do you or your organisation work with?

Only 7 respondents plan to include more languages in their work in the next 3 years, the rest are either not sure (7) or do not plan on including more languages (8).

The Wikimedia projects theoretically are able to support almost all languages as long as there is a community willing to open up a new Wikimedia project (like Wikipedia) or willing to contribute to the multilingual projects like Wikidata and Wikimedia Commons in their language which leads to the respondents having a hard time declaring which specific languages will be included in their work in the future. The languages mentioned are Estonian, Finnish, German, Latvian, Lithuanian, Danish, Cornish, Irish, Scottish Gaelic and Manx but also “Virtually any other language” and “All languages on Earth” (in response to other) and 2 respondents that work in multilingual projects have selected all the possible languages.

3.1.3. Evaluation of the Current Situation

This section of the survey sets out to assess the extent to which language technology tools and applications are used by the respondents or their organisation.

The languages for which most respondents use language technology tools (e. g., translation tools, spell/grammar checkers, web search engines, social media, language learning tools) unsurprisingly turns out to be English (14), followed by French (7), Spanish (6), German, Danish and Hungarian (5), Dutch, Irish, Swedish, Norwegian, Italian and Slovak (4). When looking at the comments of respondents we see that some respondents are stating that the

Wikimedia projects themselves can be seen as LT and thereby are used in many more languages (“Thousands of languages spoken over the world.”, “For virtually every language, 120 as of today”).

Only 12 out of 22 respondents stated that they are using language technologies for the European languages they are working with (2 said no and 3 that they don’t know). The technologies that are being used by the Wikimedia members are mainly **translation tools** (10), proofing tools (6) and search tools (6). These answers reflect the fact that the nature of the work of most of the respondents includes generating text for an online Encyclopedia (or translating words for an online dictionary). Figure 10 shows the main LT types used by the organisation.

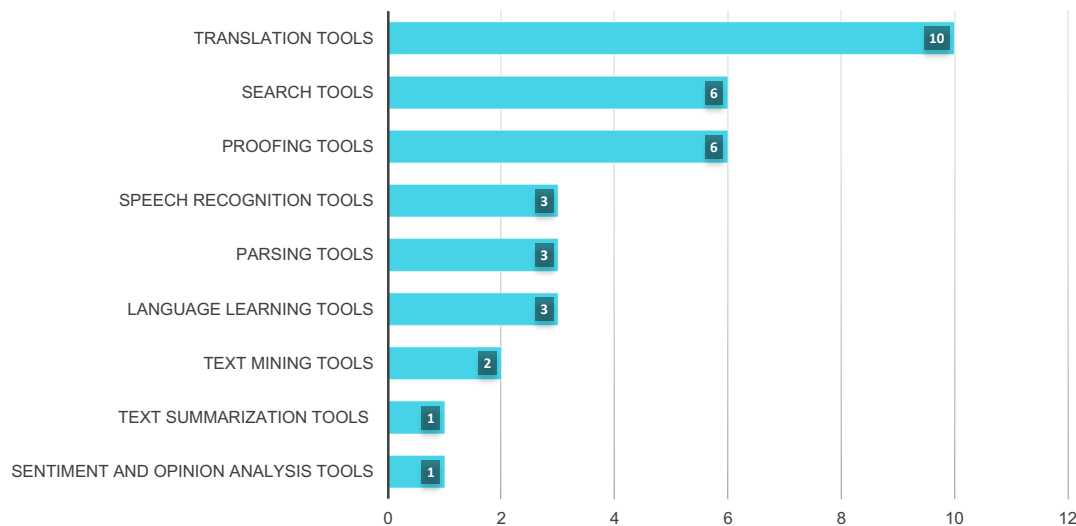


Figure 4: Which language technology tools/applications listed below do you or your organisation use with the official European language(s) you or your organisation work with?

When using translation tools, most respondents use generic translation tools freely available (8) or computer-assisted translation (6), as well as terminology management tools (1) and custom-built engines (1). The **proofing tools** that are being used are spell checkers (6), grammar checkers (5) and autocorrect tools (4). The **search tools** that were chosen are multilingual search engines (6), Generic search systems available freely on the web (e.g., Google search) (3) and Web-based question-answering systems (e.g., Stack exchange, StackOverflow, Quora, Google search) (3). The more specific search tools like domain-specific search engines (2), ontology tools (1), cross-language search engines (1) and multimedia search engines (1) do not appear to be too relevant for the respondents. Refer to Table 10 in Appendix C to see all the tools mentioned by the respondents. 8 respondents indicated that they or the organisation they work for are using language technologies to process minority, regional or lesser-used languages. Those languages were Basque (3), Catalan (3), the Erzya language, Macedonian, Breton, Mirandese, Occita and Welsh. Refer to Table 11 in Appendix C to see all the tools mentioned by the respondents.

12 (55%) out of 22 respondents have answered the question about **perceived gaps** in the technological support for their languages. While 8 respondents perceive gaps in the level of technological support, 4 do not. As our organisation is at its heart a movement for open source, open knowledge and open data, a common answer to the perceived gaps when using those languages are the lack of openly licenced resources. It has been the most frequent

response for all languages in this question. In addition to restrictive license further problems identified are (by region):

- Bulgaria: Variety of linguistic phenomena/text types covered (1)
- Danish: Amount and variety of available applications (1); Quality of the tool/application (delays in responding, difficulties with special characters, language-related errors in the output etc.) (1); Variety of linguistic phenomena/text types covered (1); The tool and all the resources (help pages etc.) only being available in English (1)
- Hungarian: Quality of the tool/application (delays in responding, difficulties with special characters, language-related errors in the output etc.) (2); Variety of linguistic phenomena/text types covered (2); Amount and variety of available applications (1)
- Icelandic: Variety of linguistic phenomena/text types covered (1)
- Irish: The tool and all the resources (help pages etc.) only being available in English (1)
- Maltese: Variety of linguistic phenomena/text types covered (2), Amount and variety of available applications; Quality of the tool/application (delays in responding, difficulties with special characters, language-related errors in the output etc.) (1); The tool and all the resources (help pages etc.) only being available in English (1)
- Slovenian: Variety of linguistic phenomena/text types covered (1)

Another problem that seems to be relevant for European languages (at least it was mentioned for Bulgarian, Danish, Hungarian, Icelandic, Maltese and Slovenian) is the variety of linguistic phenomena and text types covered. In the open answer question “In your opinion, what is going well when using these language tools?”, mentions are made to successfully using Hunspell, an open source spell checker, and occasionally finding “open source stemmers and good stop word lists”. Several respondents are mentioning that the situation has improved over the past years from “mostly unusable” to “getting better” now (although this mostly refers to the closed source tools).

When asked about the **performance of LT for their languages**, the LTs with the best performance for their respective languages were proofing tools (like spell checkers and autocorrect), translation tools (like Google translate) and search tools (e. g., Google search) (the performance of those LTs were rated by more than 6 respondents as “good” or “excellent”). For Irish and Maltese, however, the performance of proofing tools has been rated “poor” or “very poor” (same for Irish and translation tools, and Maltese and language learning tools). This might show that, while for more widely spoken languages like Danish, Hungarian, Slovenian etc. the performance of the LTs that are more commonly used (like Google search, translation tools etc.) is pretty high, this is not the case for languages with less speakers like Maltese or Irish.

When asked explicitly how they would rate the performance of LT tools in the context of the minority, regional or lesser used languages a similar picture is painted. While proofing tools, search tools and translation tools still received positive feedback (4 or more “good” or “excellent” and 2 “poor” or “very poor”), the rest of the LTs received negative feedback, if any. Again, for the other LTs (speech recognition, parsing, sentiment analysis and opinion analysis tools, text summarization, text mining, language learning) significantly less feedback was provided when compared to proofing tools, search tools and translation tools (i. e., a high number of N/A answers) which indicates that those tools are not used so much (speech recognition and language learning received only 3 answers each, all of them negative).

With respect to the frequency of use of specific LTs for minority languages, this is proven: the respondents have indicated that speech recognition, parsing, sentiment analysis and

opinion analysis tools, text summarization and text mining are “Never” used (the individual LTs received a differing number of “Never”, but the answers were all “Never”). Of the LTs outside of translation, search and proof tools, language learning tools for minority languages seem to have some relevance (with 2 respondents saying they use them “Sometimes”).

The answers relating to the questions about technological support and frequency of usage of LTs for minority, regional or lesser used languages also concentrated on search tools, translation tools and proofing tools. Proofing tools for minority languages received a favorable response for the technological support (3 “good” and 1 “excellent” answer), and similarly did translation tools (1 “very poor” and 3 “good” or “excellent”). However search tools received slightly less positive feedback for the technological support they receive (2 “very poor” or “poor” and 3 “excellent”).

With respect to the **frequency of use** of those tools, the answers seem to be correlated with the questions relating to performance. While search tools (10 “Sometimes, Frequently or Every Day”), translation tools (1 “Never” and 9 “Sometimes, Frequently or Every Day”) and proofing tools (3 “Never or Rarely” and 8 “Sometimes, Frequently or Every Day”) are more frequently used, more specific tools like sentiment analysis, text summarization and text mining (all with 5 or more answers for “Never”) are less relevant for the respondents. This seems logical when considering that the respondent would probably use the LT for article maintenance or creation, adding Lexemes or translating articles for the Wikimedia projects. For the rest of the tools, we can observe a lot of N/A answers from the respondents which indicates that the respondents do not use the rest of the LTs so frequently.

When asked about the **technological support** for the languages. Respondents were asked to rate the level of technological support based on a four-point scale (where 1 = *very poor*, 2 = *poor*, 3 = *good*, 4 = *excellent*). Unsurprisingly English enjoys the highest level of technological support from the perspectives of the respondents (between 3 and 4 rated “Excellent support”). Other languages that were well rated were Danish, German and Hungarian. While German and Danish seem to be rated more positively by respondents, the support for Hungarian language is rated less positive than for Danish and German. The other languages were not well rated or not rated (which includes the response “I do not know”) and thereby do not allow for any reasonable analysis. Figure 5 shows the mean scores given to the level of LT support per language.

3.1.4. Predictions and Visions for the Future

The survey presented the respondents with three different visions for the future when it comes to LTs in Europe. Those three visions were “In the next 10 years ...

- A ...there will be higher-quality language tools that deal with all the languages that concern me, including minority languages”,
- B ...there will be a wider range of language tools for European Languages”,
- C ...language technology tools will help prevent the loss of linguistic diversity”.

Of those three statements the second one (B) has received the most agreement from the respondents (20 “Agree” or “Strongly Agree” and 2 “Undecided”) in that this describes their vision for the future of LT best. The vision statement (C) is the one with the most extreme answers of the three: It is both the only one that has been rated with 1 “Strongly disagree” as well as the most negative statements (7 “Disagree” and “Undecided”) as well as the statement with the most “Strongly agree” answers (10). This division might be explained by the fact that it is the most ambitious and impactful of the visions and backed by the opinions of some responses to the survey and interviews that state that increasing the numbers of people speaking the language is more important than improving the technology supporting

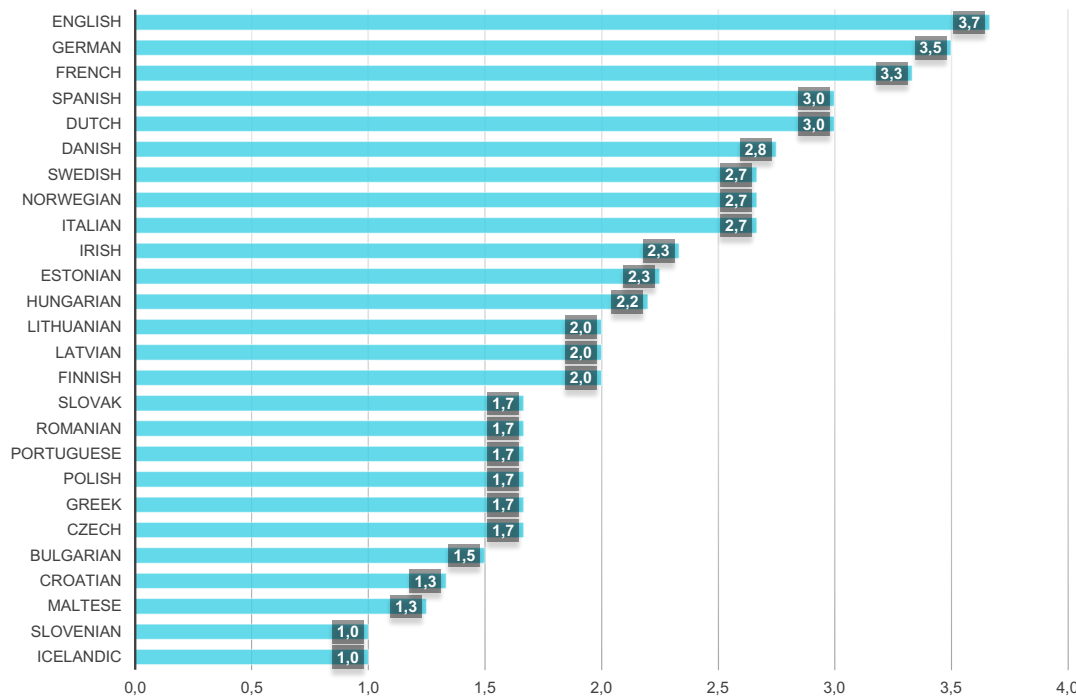


Figure 5: Please choose the option that best describes the level of language technology support for the language(s) your community or your organisation work with.

the languages. Statement (C) implies that technology will have an influence on languages surviving or not. Figure 6 shows the answers to this in more detail.

When asked about the future applications that the community members would want to see in their communities but are not available today, a lot of answers related to translation tools. In general, the respondents would like to see high quality open source translation option, translation tools that work well for regional languages (like North Frisian), real time and collaborative translation tools that allow different users to work in real time collaborative documents or Wikipedia articles, so that users can work on one document together at the same time but using different languages. Other references were made to multi-translation tools that translate to and from multiple languages at the same time, as well as translation memory support. The respondents also mentioned better voice and speech technology multiple times: a better voice recognition for all languages and voice assistants like Siri that would also be able to understand and answer questions in Welsh.

Further tools mentioned (that should be available for all languages) were sentiment analysis, grammar, style and spell checkers, as well as stemmers. More specific tools that were also part of the visions of the community members were LinguaLibre and living dictionaries for all languages as well as GPT-3 (for Russian) and a morphological analyser. As for the applications that the users would like to see for their language community, it has been stated that e.g., the interface of Google Apps (like Google Mail) should be available in lesser spoken languages (e.g., Basque) as well. Programming languages like AGDA should also be usable in all languages as well as online dating apps. More general statements were in relation to a wish for better technological support for agglutinative languages and more time for speakers to work with their languages online and contribute to the languages resources: “More learning by people than AI.”

The survey also asked about the kind of benefits there would be for those community mem-

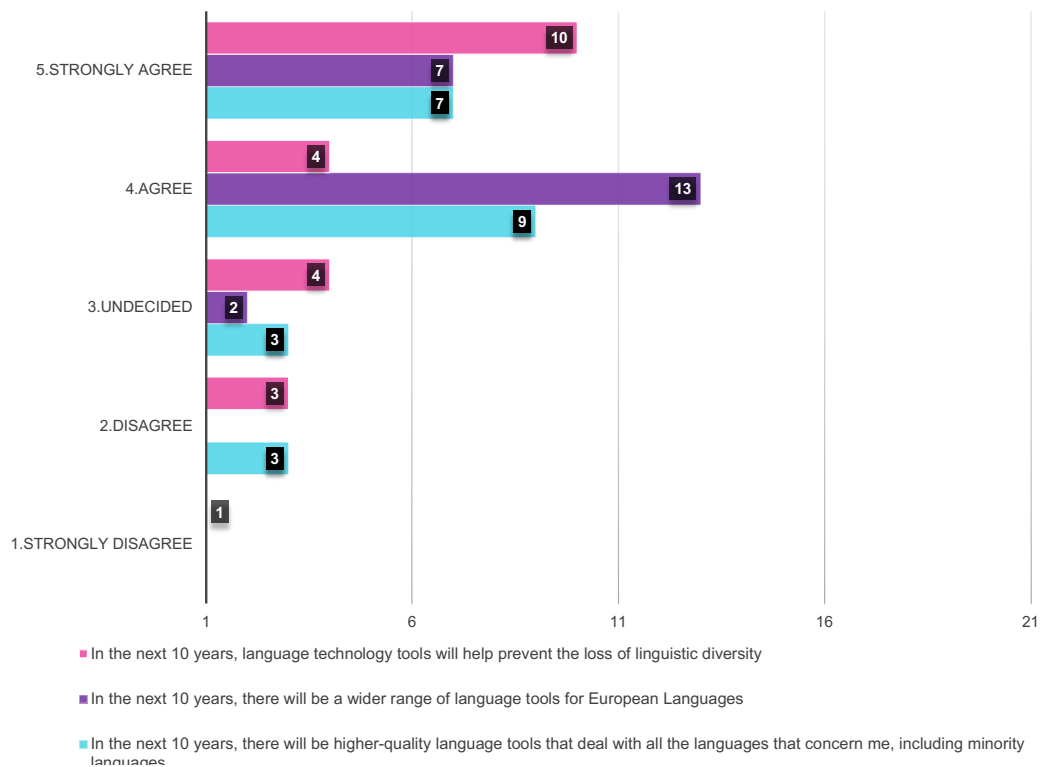


Figure 6: Please indicate the best option that describes your vision for the future of languages technology.

bers if language technologies were improved. Most respondents saw the most relevant benefit in “Preventing minority and regional languages from disappearing” (18). Also important seem to be to “Increase individuals’ exposure to these languages” (15) and “Improve the literacy for minority or regional languages” (15) as well as to “Increase engagement with social, leisure and work activities in their own languages” (14) and “Increase the number of speakers of those languages, including minority or regional languages” (13).

Less importance is being placed on the more trade and commercial orientated benefits like “Improve online trade in countries where those languages are spoken” (7) or “Improve offline trade (i. e., not e-commerce) in countries where those languages are spoken” (4). This is not surprising given that the survey was distributed in a community of volunteers.

3.2. Interview Responses

3.2.1. Gaps and Problems

During the discussions and interviews we have conducted with community members it became clear that some gaps and problems identified by the respondents and informants that related broadly to the topic of languages and language technology, were more general and basic. It was reported that, for the concerned language communities, sometimes it starts with the simple question of having basic infrastructure like a computer and internet connection available to contribute with/ to their languages online (Norge, 2021). Furthermore, hardware problems posed problems for languages such as Saami in the context of the lack of keyboard availability Norge (2021). In addition to this, other basic technological needs

that are met for those working with majority languages, are instead posing challenges to minority languages communities. These include poorly designed sites (even government sites), which break when used through minority languages. Due simply to not being considered in the design process, community members are not able to write their name in the language in certain forms because the characters cannot be processed (e.g., *k* in Skolt Saami). Respondents felt that such problems do not allow a “one-size-fits-all” solution, as was seen for the Saami languages, where it is believed that any available solution for Northern Saami could be applied to Skolt or Inari Saami. In their opinion, this would be similar to applying solutions that work for the English language to the German languages because they are “similar enough” (Interviewee 2, 2021).

Another problem stated, that is not necessarily related to language technology itself, was mentioned in the context of languages of the former Yugoslavia region: language politics and possible conflict among groups can lead to problems with contributing to languages online. This applies especially when trying to find standards for the language, which then will be applied in the Wikimedia projects or other resources for LT (CEESession, 2021). It was stated in the context of the CEE meeting that being rigid about a certain standard of language and trying to have this standard reflected in the Wikimedia content is a widespread phenomena in the Wikimedia movement. As consequence, there can be frustration for new users who don’t have enough standing in the communities to discuss applying certain language standards (CEESession, 2021). It was highlighted that many pluricentric languages (e.g., Serbo-Croatian) do not have a central organization that defines language standards like there are for some major pluricentric languages (like French or German) (CEESession, 2021). The lack of a written language was also mentioned for the Saami languages. In general it was stated that the technological support for standard languages is usually on a higher level than those for local variants. During the CEE Meetup it was also discussed that machine translation is a language technology that works especially poorly for Slavic languages (CEESession, 2021).

3.2.2. Lack of language experts in the Wikimedia Community

As the Wikimedia movement is built by volunteers and shaped by volunteer work, the volunteer community is at heart of all the things we as Wikimedia do. It is important to therefore focus also on challenges faced by the communities working with those languages online, or in revitalization projects or with language technology in their fields.

In general, it was stated (similar to the open answers in the survey) that there is a need for more speakers of under-resourced languages that can contribute to and use the language technologies mentioned. The Wikimedia projects serve as an important base for many language technology applications and rely heavily on volunteer work. However not all language community members have the same privilege to work for free or little money on these kind of projects. This has been mentioned both in international discussions with a focus wider than just European languages, as well as in the context of minority language communities in Europe (Interviewee 2, 2021). One suggested solution would be to find ways of financially supporting those members of the small language communities in contributing to our projects (Norge, 2021). It is important that actual speakers of the language deal with the project and technologies, as non-fluent speakers, learners or non-speakers could cause damage when trying to working with these resources⁸. This results in the need for expert editors to correct inaccurate content (Interviewee 2, 2021). In order to have a successful and active Wikimedia project, not only are speakers of the language needed, but also technically savvy contributors (Norge, 2021).

⁸ <https://www.theguardian.com/uk-news/2020/aug/26/shock-an-aw-us-teenager-wrote-huge-slice-of-scots-wikipedia>

3.2.3. Recommendations

During the course of our discussion rounds and through the analysis of past events we have compiled recommendations that were made by Wikimedia community members.

The interface of tools has been mentioned as one area for improvement. Translation tools require further user interface development to make them more user-friendly (Interviewee 1, 2021). Localisation of tool interfaces should be expanded to include more languages. One such interface that was highlighted was that of MediaWiki, where language communities can set up their own Wikimedia projects using their languages (WDConSession, 2021). Another example mentioned was the interfaces of EU and EC websites that don't offer Northern Saami or Skolt Saami translations (Interviewee 2, 2021). Another recommendation has been to update CLDR specifically for under-resourced languages, as this information is being used by major companies whose products have a big influence on many people's lives (Interviewee 2, 2021).

Further resources and tools that have been recommended for development have been: "dedicated keyboard layouts specific to that language, recording digital stories in one's own language, promoting subtitling initiatives, streaming online using free software tools, collecting publicly available linguistic data from the web, creating a Wikimedia project in minor languages, sponsoring initiatives to increase the size and quality of content of the Wikimedia projects in that language, develop smartphone apps and video games." ("Diversity", 2019). Giellatekno⁹, a center for Saami language technology, was mentioned as a best practice where LT is produced for the Saami languages (online dictionaries, spell checkers etc.) and which should be supported (Interviewee 2, 2021).

Funding initiatives that are dealing with small and lesser used languages were also mentioned. It was recommended that under-resourced languages, as well as national libraries of those languages should be a priority for public funding such as grants offered by the European Commission (Interviewee 1, 2021).

The funding situation for minority and under-resourced languages (e.g., the Saami languages) has been described as "meaningless", also because funds do not allow for a long term perspective for the language communities (Interviewee 2, 2021).

Another recommendation revolved around the need for greater awareness of existing resources and projects. One specific example of this is the range of Wiktionary projects, which have a high quality and are not just dictionaries but have ethnology, pronunciation and grammar (especially in the CEE region, like Polish Wiktionary). There are Wiktionaries available for a many languages in the CEE region, yet there is no real awareness of this in the fields of interpretation or translation (especially in professional settings) CEESession (2021).

As of now a lot of lesser used, medium or smaller languages have weak links to each other in terms of resources. This means that, for example, translating directly from Swedish to Hungarian can be a challenge. A suggestion to help with this is for the EU to focus developing tools and resources for major languages (such as Turkish, Arabic, Russian, German) that can work as pivot languages for lesser-spoken languages. Learning a new language, it is hard to find resources for Swedish-Hungarian, but resources for German-Hungarian and German-Swedish exist and someone learning Hungarian as a Swedish speaker could use the existing German resources – given German language skills. If this approach would actually help smaller, medium or lesser used languages surviving in the long run could be questioned though (CEESession, 2021).

As mentioned before, a big topic for community members in the Wikimedia community is open access to language resources and tools. Naturally, a reoccurring remark during our conversations was the lack of available resources due to restrictive resources. This was mentioned for official dictionaries (and similar resources) (WDConSession, 2021), as well as lan-

⁹ <https://giellatekno.uit.no/index.eng.html>

guage tools such as spell checkers for Saami languages¹⁰ (Norge, 2021). For some languages though, the questions around available open resources is even more basic and has to focus on learning material (like school books that partially “are 10 years behind the ones in the majority languages”) (Interviewee 2, 2021). Availability of CC0 material makes quite a difference (which e.g., is available for Dutch) (WLDH). Wikidata as a valuable resource is also mentioned: It is recommended to include Wikidata in EU resources more (WDConSession, 2021).¹¹ Another best practice mentioned for digital language equality for minority languages has been to create general ontologies like the existing ones for bigger languages (e.g., YSO for Finnish) also for smaller and minority languages (Interviewee 2, 2021).

3.2.4. Wikidata, Lexemes and Abstract Wikipedia

In the context of supporting small and minority languages, Wikidata and its potential to contribute to the survival of languages has been discussed throughout the Wikimedia movement (WikidataCon, Celtic Knot) and been mentioned as an easy way for small language communities to contribute themselves to the digital survival of their languages through the use of Wikimedia projects and language technologies (Interviewee 1, 2021).

Wikidata is a free, collaborative, multilingual knowledge base with a focus on verifiability. It collects structured data to provide support for Wikipedia, the other wikis in the Wikimedia movement, and anyone in the world with a need for general-purpose structured data. Wikidata is based on the Wikibase software and provides data, an ontology and links to other databases. Wikidata’s data constitutes the basis for a wide variety of applications and services both inside and outside the Wikimedia movement. It is an increasingly important building block for much of the technology we use every day. Wikidata has just turned nine years old and is thriving more than ever. Wikidata describes almost 100 Million concepts, and crossed the 1.5-billion edit mark in 2021. A full 72% of Wikipedia articles use Wikidata for infobox content, auto-categorization, flagging maintenance work and other support functionality, not to mention site links, which are used in 97% of all Wikipedia articles. The Wikidata Query Service, a SPARQL endpoint for querying Wikidata’s graph, sees 11 million queries per day. Wikidata has also successfully expanded into the area of lexicographical data, forming the basis for new initiatives such as Abstract Wikipedia.” (WMDE, 2021).

This lexicographical data is structured information about words (and their forms and senses) that is stored in Wikidata and machine-readable.¹² The Lexemes in Wikidata are used to describe words whereas the Items in Wikidata are used to describe concepts. As Items and Lexemes can be linked within Wikidata, machines can not only understand the word, but also the meaning of the word (the lexicographical and the semantic meaning). This possibility to edit and add Lexemes to the Linked and Open Data Web via Wikidata has been introduced in 2018. To this day (25/11/2021) 605.158 Lexemes have been added by the Wikidata community to Wikidata in 799 languages¹³. The languages with the most Lexemes as of now are Russian (101.322), Estonian (83.208), English (71.293) and Malayala (62.954) (see Table 3 in the appendix for an overview of numbers of Lexemes for the ELE language selection). With a comprehensive corpus of lexicographical data in Wikidata, there are a number of visionary applications that can be build on top of and with the help of this data.

- Creation of comprehensive dictionaries and automated translations from small or minority languages to small or minority languages: Because of the linked nature of Lexemes, words could be connected from any added language to any other added language and direct translation from e.g., North Frisian to Occitan would be possible.

¹⁰ <https://giellalt.github.io/LanguageModels.html>

¹¹ <https://ec.europa.eu/jrc/en/digcomp/digital-competence-framework>

¹² https://www.wikidata.org/wiki/Wikidata:Lexicographical_data

¹³ <https://ordia.toolforge.org/statistics/>

- Easier creation of specialized dictionaries: Due to the linked and queryable nature of Wikidata and its Lexemes and Items, specialized lists that only include information for certain topics (like art, sports, science etc.) can be created in an automated way
- Language Learning Tools: Similar to specialized dictionaries, Wikidata and its Lexemes and Items can help to create specialized word lists for lessons on specific topics more easily
- Furthermore this kind information can enable and facilitate much more in depth and detailed linguistic research about connections between languages and words and how they evolve over time
- Text analysis: Machine-readable lexicographical data is one important building block for analysing the content of texts. This enables sentiment analysis, part of speech tagging and named entity recognition for example.
- Automated text generation from any language added to Wikidata and the Lexemes: The lexicographical data from Wikidata will be the basis for Abstract Wikipedia.¹⁴

Abstract Wikipedia is a new addition to the Wikimedia projects and is envisioned to provide the opportunity for natural language text creation for any language by combining a catalogue of (community created and maintained) functions from Wikifunctions with the all-purpose and lexicographical data from Wikidata. The texts that will be produced with this community-driven project will then be available for any language version Wikipedia – no matter how big or small a language community might be.

3.2.5. Initiatives for building audio corpus for languages/ Oral Knowledge

The Wikimedia movement has historically relied heavily on written knowledge to create and verify its content. We have, however, recognized that in order to include knowledge from all languages, we need to understand that part of this knowledge is embedded in a tradition of oral knowledge. Projects and initiatives like Wikitongues tries to capture languages in this tradition and to “ensure every person has the tools to preserve, promote, and pass their languages on to the next generation”. In addition to distributing accessible frameworks for language preservation, we’re building a public archive of oral histories in every language in the world.¹⁵

Lingua Libre is another initiative that has provided a tool for building an audiovisual corpus for languages throughout the world that is available under a free license. This online solution allows speakers to record pronunciations of words in their language. The recordings are queryable and part of the Linked Open Data web (Movement, 2021c). As for the languages that are being covered by the ELE consortium, not all, but a growing number of European languages are represented in Lingua Libre (see Table 4). The ELE languages with the largest amount of data and individual audio recordings are French (231.235), Polish (54.846), Romanian (19.400) and English (19.353). Lingua Libre was developed in the midst of the French community, which explains the high number of Items for French and also French minority languages being represented in the data more so than other European languages (Occitan having a bigger corpus than Swedish for example). This leads us to the conclusion that if the responsibility and resources for the development of language technologies and the production or corpus is put in the hands of the affected communities, the output for the affected languages are higher.

¹⁴ https://meta.wikimedia.org/wiki/Abstract_Wikipedia

¹⁵ <https://meta.wikimedia.org/wiki/Wikitongues>

4. Conclusions

The ELE project has consulted Wikimedia Deutschland to share the perspective of Wikimedia community members on Digital Language Equality in terms of language technologies and their experiences working with their languages in an online environment. The various Wikimedia projects have strong connections to languages and the Wiki communities have made a valuable contribution to keeping languages (smaller, minority, regional, lesser used languages) alive in an online environment. The ELE languages are being represented in the Wikimedia movement both by official chapters, by less formal user groups, groups of communities and individual volunteers working with those languages in the Wikimedia projects.

Through the survey, the discussion rounds and the interviews, we managed to capture the voices of both representatives of the Wikimedia chapters and user groups and of volunteers working in the Wikimedia projects. The 22 respondents to the survey represented users/speakers of language technologies for a range of languages including: English, French, Spanish, Danish, German, Hungarian, Dutch, Irish, Italian, Norwegian and Slovak. There is a bias, however, in terms of the proportion of these language communities that have contributed to this report. This is mainly due to our dissemination approaches; discussion rounds we organised with the CEE community (e. g., Hungarian, Slovak) and the announcement of the study during the course of the Wikimedia language conference – Arctic Knot (e. g., Saami languages) – reached these communities more easily. The answers given in those discussion rounds were more representative of those few language communities. However, we also managed to include voices from speakers of Irish, Welsh and Maltese.

This survey has revealed that, in terms of language technology, the Wikimedia communities are mainly using translation tools, proofing tools and search tools. This holds true for both widely-spoken major languages and minority or lesser-spoken languages. As we are dealing with a community that is producing or maintaining encyclopedic texts, adding content to online projects, translating articles or adding lexicographical information to online projects, the higher prevalence of use across those three types of LTs is not surprising. When looking at the performance of LTs, the ratings are also much more positive for the translation tools, search tools and proofing tools. Other LT tools listed in the survey are rarely used by the respondents and as such not much feedback was offered. The same picture is being painted when asked about the technological support. A big part of the positive feedback is concentrated on major languages like English and German, whereas technological support for languages like Hungarian, Maltese and Irish received less favourable feedback.

The survey and discussion groups revealed that the **main gaps and issues** related to the lack of open-source resources – which holds true for all of the languages considered. Furthermore, it has been discovered that the challenges of working with some minority languages like the Saami languages online are not specifically LT-related, but instead relate to basic problems like lack of internet connection, computers or specific keyboards for the languages exist for smaller language communities. One aspect that has been mentioned throughout discussion rounds, interviews and survey is that the lack of LTs or LT resources is not the most pressing issue for small, regional, minority or under-resourced language communities, but the lack of speakers of those languages that could contribute to the digital survival of the languages in the first place.

Summary of problems and major gaps for Language Technology

- Lack of open-source resources, which is especially relevant for minority languages (language learning materials, school books, open-source dictionaries, translations resources, stop words, stemmers, written documents, audio data or spell checkers)
- Lack of translations of interfaces (e. g., Google Apps) or websites for minority languages

or poor quality and non-functioning of those interfaces for minority languages (e. g., EC websites not being available in minority languages)

- Language tools and the documentation about them only being available in English
- Variety of linguistic phenomena and texts (e. g., Bulgarian, Danish, Hungarian, Icelandic, Maltese and Slovenian)
- Poor quality of voice recognition for regional and minority languages (e. g., Welsh)

Non-Language Technology related problems

- Lack of language experts of minority languages contributing to their languages online, e. g., in the language versions Wikipedia of their languages
- Lack of long-term funding for projects and institutions (e. g., libraries) working with regional and minority languages
- Lack of keyboards for specifics of minority languages (e. g., Saami languages)
- Lack of certain infrastructures for minority language communities (good internet connection, computers)
- Lack of standardization for some pluricentric language

The **visions** we have discussed in this report are for innovative and open source translation tools for all languages, especially for under-resourced languages: multilingual translation tools (translating in multiple languages at once) or real-time collaborative translation tools that allow speakers of different languages to work together on one text. Wikidata and the lexicographical data stored there is a visionary project pushed by the Wikimedia movement that provides opportunities for small communities to contribute to the digital survival of their languages. Including oral knowledge and making sure contributing to audio content and building audio corpus is another vision the Wikimedia movement and volunteers have worked towards, with the establishment of LinguaLibre for example.

Visions for Language Technologies in the future

- Translations tools working as good for regional languages like North Frisian as for English
- Extensive information and resources about all regional and minority languages in a linked open data environment, e. g., in the form of lexicographical data on Wikidata
- Online dating apps available for all languages
- Tools translating text and document to and from multiple languages
- Programming Languages and programming environments (e. g., MediaWiki and AGDA) available for languages other than English
- Real-time and collaborative translations tool to work collaboratively on document in several languages
- Inclusion of non-written languages in the digital sphere

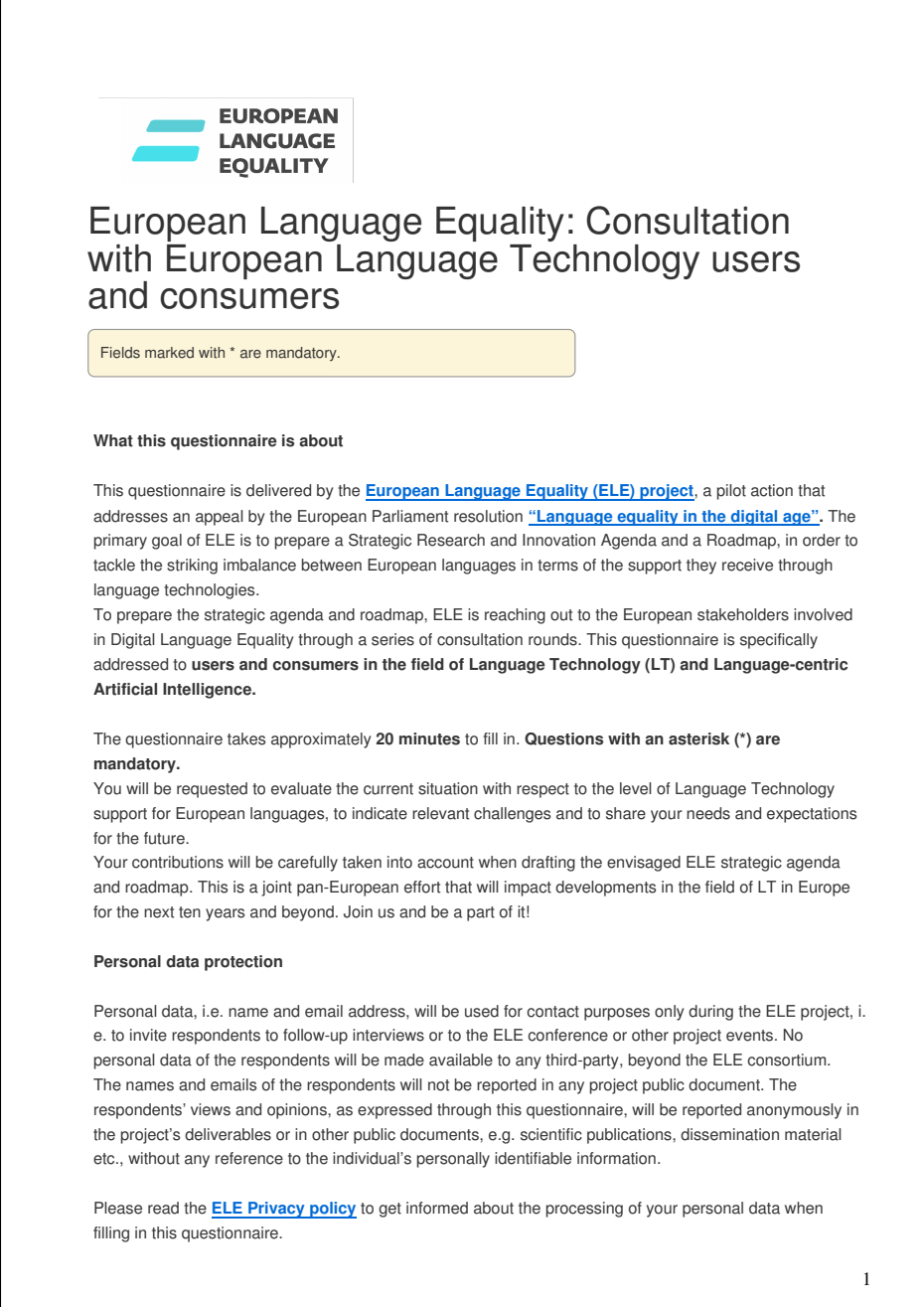
We hope that the efforts of all the language communities and Wikimedia communities will contribute to a situation where all languages can enjoy equal representation in the digital sphere.


References

- CEESession. Discussion Round conducted by Maria Heuschkel at CEE conference 2021 “Get together – Partnering around Languages in Europe!”, 2021. Session on November 6 <https://www.youtube.com/watch?v=PY9MamTE91U&t=330s>.
- Wikimedia Strategy Working Group “Diversity”. Establishing partnerships in order to represent and protect worlds’ cultural diversity, September 2019. https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20/Recommendations/Iteration_2/Diversity/6.
- Maria Heuschkel. Talk at the Arctic Knot Conference 2021, 2021. https://meta.wikimedia.org/wiki/Arctic_Knot_Conference_2021/Submissions/ELE_lightning_talk.
- Interviewee 1. Interview 1 conducted by Maria Heuschkel, 2021. Questionnaire from November 16 2021.
- Interviewee 2. Interview 2 conducted by Maria Heuschkel, 2021. Questionnaire from December 30 2021.
- Wikimedia Movement. Statistics on Wikipedia, 2021a. retrieved November 2021 from https://commons.wikimedia.org/wiki/Data:Wikipedia_statistics/meta.tab.
- Wikimedia Movement. List of Wikimedia projects by size, 2021b. retrieved November 2021 from https://meta.wikimedia.org/wiki/List_of_Wikimedia_projects_by_size.
- Wikimedia Movement. Lingua Libre, 2021c. https://meta.wikimedia.org/wiki/Lingua_Libre.
- Wikimedia Norge. Workshop for a future Wikimedia Language Diversity hub at Wikimania 2021, August 2021. <https://etherpad.wikimedia.org/p/Wikimania2021-LangHub-2>.
- WDConSession. Discussion Round conducted by Maria Heuschkel at Wikidata Conference 2021 “Digital Language Equality for all Languages in Europe – The ELE project”, 2021. Session on October 30 <https://pretalx.com/wdcon21/talk/NHRWTH/>.
- WikimediaMovement. Wikimedia movement affiliates, 2021. retrieved November 2021 from https://meta.wikimedia.org/wiki/Wikimedia_movement_affiliates#chapters.
- WMDE. Linked Open Data Strategy - Wikidata, 2021. <https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Wikidata>.

A. LT users and consumers survey

Figures 7 to 24 show the complete LT research and developers survey.



 **EUROPEAN
LANGUAGE
EQUALITY**

European Language Equality: Consultation with European Language Technology users and consumers

Fields marked with * are mandatory.

What this questionnaire is about

This questionnaire is delivered by the [European Language Equality \(ELE\) project](#), a pilot action that addresses an appeal by the European Parliament resolution "[Language equality in the digital age](#)". The primary goal of ELE is to prepare a Strategic Research and Innovation Agenda and a Roadmap, in order to tackle the striking imbalance between European languages in terms of the support they receive through language technologies.

To prepare the strategic agenda and roadmap, ELE is reaching out to the European stakeholders involved in Digital Language Equality through a series of consultation rounds. This questionnaire is specifically addressed to **users and consumers in the field of Language Technology (LT) and Language-centric Artificial Intelligence**.

The questionnaire takes approximately **20 minutes** to fill in. **Questions with an asterisk (*) are mandatory.**

You will be requested to evaluate the current situation with respect to the level of Language Technology support for European languages, to indicate relevant challenges and to share your needs and expectations for the future.

Your contributions will be carefully taken into account when drafting the envisaged ELE strategic agenda and roadmap. This is a joint pan-European effort that will impact developments in the field of LT in Europe for the next ten years and beyond. Join us and be a part of it!

Personal data protection

Personal data, i.e. name and email address, will be used for contact purposes only during the ELE project, i. e. to invite respondents to follow-up interviews or to the ELE conference or other project events. No personal data of the respondents will be made available to any third-party, beyond the ELE consortium. The names and emails of the respondents will not be reported in any project public document. The respondents' views and opinions, as expressed through this questionnaire, will be reported anonymously in the project's deliverables or in other public documents, e.g. scientific publications, dissemination material etc., without any reference to the individual's personally identifiable information.

Please read the [ELE Privacy policy](#) to get informed about the processing of your personal data when filling in this questionnaire.

1

Figure 7: Full survey as published (page 1/18)

Introduce yourself and your organisation

*** In which country are you based?**

☐ Austria ☐ Germany ☐ Poland
☐ Belgium ☐ Greece ☐ Portugal
☐ Bulgaria ☐ Hungary ☐ Romania
☐ Croatia ☐ Ireland ☐ Slovak Republic
☐ Cyprus ☐ Italy ☐ Slovenia
☐ Czechia ☐ Latvia ☐ Spain
☐ Denmark ☐ Lithuania ☐ Sweden
☐ Estonia ☐ Luxembourg ☐ Other
☐ Finland ☐ Malta
☐ France ☐ Netherlands

*** If "other", please specify.**

*** Which community are you representing? (e.g. Wikidata, Italian Wikipedia, User Groups etc.)**

*** What is the name of the organisation/representative body you work for or the name of the project you are active in?**

*** Which of the following best describes the type of organisation you work for or the community you are representing?**

☐ Professional associations
☐ Government department/unit
☐ SME
☐ Large Enterprise
☐ Independent contractor/ consultant
☐ Education/research
☐ N/A
☐ Other

2

Figure 8: Full survey as published (page 2/18)

* If "other", please specify.

* What is the size of the organisation or the project you represent/ work for/ are active in (e.g. total number of full-time employees or number of active volunteers per month)?

☐ 1-10
☐ 11-100
☐ 101-500
☐ 501-5000
☐ More than 5000
☐ N/A

* What is your main role in the organisation body you work for or the project you are active in? (if you are self-employed or if you are not employed, please specify)

If applicable: What is your organisation's estimated annual revenue in Euro?

Language Coverage

* For which language (s) you, your community or your organisation use language technology tools (e.g. Translation tools, Spell/grammar checkers, web search engines, social media, language learning tools)?

<input type="checkbox"/> Bulgarian	<input type="checkbox"/> German	<input type="checkbox"/> Norwegian
<input type="checkbox"/> Croatian	<input type="checkbox"/> Greek	<input type="checkbox"/> Polish
<input type="checkbox"/> Czech	<input type="checkbox"/> Hungarian	<input type="checkbox"/> Portuguese
<input type="checkbox"/> Danish	<input type="checkbox"/> Icelandic	<input type="checkbox"/> Romanian
<input type="checkbox"/> Dutch	<input type="checkbox"/> Irish	<input type="checkbox"/> Slovak
<input type="checkbox"/> English	<input type="checkbox"/> Italian	<input type="checkbox"/> Slovenian
<input type="checkbox"/> Estonian	<input type="checkbox"/> Latvian	<input type="checkbox"/> Spanish
<input type="checkbox"/> Finnish	<input type="checkbox"/> Lithuanian	<input type="checkbox"/> Swedish
<input type="checkbox"/> French	<input type="checkbox"/> Maltese	<input type="checkbox"/> Other

* If "other", please specify.

3

Figure 9: Full survey as published (page 3/18)

*** Do you, your community or your organisation plan to include more languages in your work in the next 3 years?**

☐ Yes
☐ No
☐ Not sure

*** Which language(s)?**

<input type="checkbox"/> Bulgarian	<input type="checkbox"/> German	<input type="checkbox"/> Norwegian
<input type="checkbox"/> Croatian	<input type="checkbox"/> Greek	<input type="checkbox"/> Polish
<input type="checkbox"/> Czech	<input type="checkbox"/> Hungarian	<input type="checkbox"/> Portuguese
<input type="checkbox"/> Danish	<input type="checkbox"/> Icelandic	<input type="checkbox"/> Romanian
<input type="checkbox"/> Dutch	<input type="checkbox"/> Irish	<input type="checkbox"/> Slovak
<input type="checkbox"/> English	<input type="checkbox"/> Italian	<input type="checkbox"/> Slovenian
<input type="checkbox"/> Estonian	<input type="checkbox"/> Latvian	<input type="checkbox"/> Spanish
<input type="checkbox"/> Finnish	<input type="checkbox"/> Lithuanian	<input type="checkbox"/> Swedish
<input type="checkbox"/> French	<input type="checkbox"/> Maltese	<input type="checkbox"/> Other

*** If "other", please specify.**

*** Do you, your community or your organisation use language technologies to process any minority /regional/lesser-used language(s) not included in the list of EU languages provided above?**

Minority languages/regional/lesser-used languages are languages that are traditionally used within a given territory of a state by nationals of that state who form a group numerically smaller than the rest of the state's population and [are] different from the official language(s) of that state" (Council of Europe, 1992, p. 2)

☐ Yes
☐ No

*** Which minority/regional/lesser-used language(s)?**

[Evaluation of the current situation](#)

4

Figure 10: Full survey as published (page 4/18)

*** Are there language technology tools/applications available for the European language(s) you, your community or your organisation deal with?**

☐ Yes
☐ No
☐ I do not know

*** Which tools/applications do you use with these languages?**

For examples of these types of tools/applications, click on boxes and select as many as apply.

<input type="checkbox"/> Proofing tools	<input type="checkbox"/> Sentiment and opinion analysis tools
<input type="checkbox"/> Translation tools	<input type="checkbox"/> Text summarization tools (e.g. Quilbot AI)
<input type="checkbox"/> Speech recognition tools	<input type="checkbox"/> Text mining tools (e.g. IBM Watson)
<input type="checkbox"/> Parsing tools	<input type="checkbox"/> Language learning tools
<input type="checkbox"/> Search tools	<input type="checkbox"/> Other

*** Proofing tools**

Please, select as many as apply.

☐ Spell checkers
☐ Grammar checkers
☐ Autocorrect tools

*** Translation tools**

Please, select as many as apply.

☐ Computer-assisted translation tools (e.g. translation memories)
☐ Terminology management applications
☐ Generic translation tools freely available on the web (e.g. Google Translate)
☐ Custom-built translation engines

*** Speech recognition tools**

Please, select as many as apply.

☐ Voice user interfaces (e.g. Siri, native android, native iOS, smart speakers [Google home, Alexa, ...], Bose Headphones, Adobe Acrobat reader, Amazon Polly, Chromevox, Wordreference)
☐ Text-to-speech systems (i.e. systems that turn text into speech for reading texts out loud (e.g. Amazon Polly, Adobe Acrobat reader)

*** Parsing tools**

Please, select as many as apply.

☐ Dependency or constituency parsing systems to automatically analyse the syntax of textual or spoken data (e.g. Stanford NLP's CoreNLP java framework, Stanford NLP Stanza, AllenNLP parsing, UDPipe, MaChAmp)
☐ Part-of-speech taggers of any type (e.g. NLTK python library, NLPdotnet)

*** Search tools**

Please, select as many as apply.

☐ Web-based question-answering systems (e.g. Stack exchange, StackOverflow, Quora, Google search)
☐ Ontology tools for extracting the corresponding domain's terms and the relationships between the concepts that these terms represent in a text (e.g. Robot tool)

5

Figure 11: Full survey as published (page 5/18)

☐ Generic search systems freely on the web (e.g. Google search)

☐ Customer-build search engines (e.g. organisations or vendors create search engines themselves)

☐ Domain-specific search engines (focusing on domain-specific topics, e.g. PubMed, Copernic, CC search)

☐ Multilingual search engines (e.g. Google, Wikipedia)

☐ Cross-language search engines (e.g. eBay, Aliexpress)

☐ Language-focused search engines (e.g. Baidu)

☐ Multimedia search engines (e.g. plantnet, or applications like 'Snooth')

☐ Private search engines (e.g. Search Encrypt and OneSearch, use different encryption methods to keep your query private)

*** Language learning tools**

Please, select as many as apply.

☐ Computer-assisted language learning tools (e.g. Duolingo, FluentU, SKELL)

☐ Web-based thesaurus tools (help users to find synonyms of words)

☐ Intelligent systems to aid and assess reading comprehension (e.g. Whooo's Reading, Storia)

☐ Web-based translation search engines (e.g. Linguee)

*** If "other" tool(s), please specify.**

*** Do you perceive gaps in technological support for the EU language(s) you work with?**

☐ Yes

☐ No

6

Figure 12: Full survey as published (page 6/18)

What are the main problems (or poor results) you observe when using these language tools?
Please, select as many gaps and languages as apply.

	Amount and variety of available applications	Quality of the tool /application (delays in responding, difficulties with special characters, language-related errors in the output etc.)	Variety of linguistic phenomena /text types covered	Adaptability to systems (e.g. adaptability to iOS system)	The tool and all the resources (help pages etc.) only being available in English	Unavailability of resources in the language because of restrictive licences	Other
Bulgarian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Croatian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Czech	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Danish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dutch	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
English	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Estonian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Finnish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
French	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
German	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Greek	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

7

Figure 13: Full survey as published (page 7/18)

Hungarian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Icelandic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Irish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Italian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Latvian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lithuanian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maltese	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Norwegian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Polish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Portuguese	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Romanian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Slovak	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Slovenian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spanish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Swedish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

8

Figure 14: Full survey as published (page 8/18)

* If "other", please specify.

In your opinion, what is going well when using these language tools?

In general terms, how do you evaluate the performance of the tools you use for the language(s) you work with?
Please evaluate based on a four-point scale.
 Please, evaluate as many tools as apply. If you do not know one or more tools, please select non-applicable (N/A).

	1. Very poor	2. Poor	3. Good	4. Excellent	5. N/A
Proofing tools (e.g. Spell checkers, Autocorrect)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Translation tools (e.g. Google Translate)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speech recognition tools (e.g. Siri, Alexa)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parsing (e.g. PoS taggers)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search tools (e.g. Google search)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sentiment analysis and opinion analysis tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text summarization (e.g. Quillbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text mining (e.g. IBM Watson)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Language learning (e.g. Duolingo, thesaurus, bilingual dictionaries)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

* If "other", please specify.

Please choose the option that best describes the level of language technology support for the language(s) your community or your organisation work with.
 Please, choose as many languages as apply.

9

Figure 15: Full survey as published (page 9/18)

	1. No support	2. Poor support	3. Good support	4. Excellent support	5. I do not know
Bulgarian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Croatian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Czech	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Danish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dutch	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
English	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Estonian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Finnish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
French	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
German	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Greek	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hungarian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Icelandic	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Irish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Italian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Latvian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lithuanian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Maltese	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Norwegian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Polish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Portuguese	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Romanian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Slovak	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Slovenian	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spanish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Swedish	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Please indicate based on a five-point scale how frequently you, your community or your organisation use the language technology tools/applications listed below for the languages you work with.

Please, select as many tools as apply.

10

Figure 16: Full survey as published (page 10/18)

	1. Never	2. Rarely	3. Sometimes	4. Frequently	5. Every day
Proofing tools (e.g. Spell checkers, Autocorrect)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Translation tools (e.g. Google Translate)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speech recognition tools (e.g. Siri, Alexa)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parsing (e.g. PoS taggers)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search tools (e.g. Google search)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sentiment analysis and opinion analysis tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text summarization (e.g. Quillbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text mining (e.g. IBM Watson)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Language learning (e.g. Duolingo, thesaurus, bilingual dictionaries)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

* If "other" tool(s), please specify.

Please indicate for which language(s) you or your organisation use the language technology tools /applications listed below.

Please, select as many tools and languages as apply.

	Proofing tools (e.g. Spell checkers, grammar checkers)	Translation tools (e.g. Google Translate)	Speech Recognition tools (e.g. Siri, Alexa)	Search tools (e.g. Google search, Wikipedia)
Bulgarian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Croatian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Czech	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Danish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Dutch	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
English	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Estonian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Finnish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 17: Full survey as published (page 11/18)

French	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
German	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Greek	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Hungarian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Icelandic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Irish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Italian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Latvian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Lithuanian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Maltese	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Norwegian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Polish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Portuguese	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Romanian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Slovak	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Slovenian	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Spanish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Swedish	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Other	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

* If "other" language(s), please specify.

Are there language technology tools/applications available for the minority/regional/ lesser-used language(s) you, your community or your organisation deal with?

☐ Yes
☐ No
☐ I do not know

*** Which tools/applications do you use with these minority/regional/lesser-used languages?**

For more examples of these types of tools, click on the boxes and select as many tools as apply.

☐ Proofing tools

☐ Search tools

☐ Language learning tools

☐ Translation tools

☐ Sentiment and opinion analysis tools

☐ Other

12

Figure 18: Full survey as published (page 12/18)

☐ Speech recognition tools
 ☐ Text summarization tools (e.g. Quilbot AI)

☐ Parsing tools
 ☐ Text mining tools (e.g. IBM Watson)

*** Proofing tools**

Select as many as apply.

☐ Spell checkers
 ☐ Grammar checkers
 ☐ Autocorrect

*** Translation tools**

Select as many as apply.

☐ Computer-assisted translation tools (e.g. translation memories)
 ☐ Terminology management applications
 ☐ Generic translation tools freely available on the web (e.g. Google Translate)
 ☐ Custom-built translation engines

*** Speech recognition/synthesis tools**

Select as many as apply.

☐ Voice user interfaces (e.g. Siri, native android, native iOS, smart speakers [Google home, Alexa, ...], Bose Headphones, Adobe Acrobat reader, Amazon Polly, Chromevox, Wordreference)
 ☐ Text-to-speech systems (i.e. systems that turn text into speech or for reading text out loud (e.g. Amazon Polly, Adobe Acrobat reader))

*** Parsing tools**

Please, select as many as apply.

☐ Dependency or constituency parsing systems to automatically analyse the syntax of textual or spoken data (e.g. Stanford NLP's CoreNLP java framework, Stanford NLP Stanza, AllenNLP parsing, UDPipe, MaChAmp)
 ☐ Part-of-speech taggers of any type (e.g. NLTK python library, NLPdotnet)

*** Search tools**

Please, select as many as apply.

☐ Web-based question-answering systems (e.g Stack exchange, StackOverflow, Quora, Google search)
 ☐ Ontology tools for extracting the corresponding domain's terms and the relationships between the concepts that these terms represent in a corpus (e.g. Robot tool)
 ☐ Generic search systems freely on the web (e.g. Google search)
 ☐ Customer-build search engines (e.g organisations or vendors create search engines themselves)
 ☐ Domain-specific search engines (focusing on domain-specific topics, e.g. PubMed, Copernic, CC search)
 ☐ Multilingual search engines (e.g. Google, Wikipedia)
 ☐ Cross-language search engines (e.g. eBay, Aliexpress)
 ☐ Language-focused search engines (e.g. Baidu)
 ☐ Multimedia search engines (e.g. plantnet, or applications like 'Snooth')
 ☐ Private search engines (e.g. Search Encrypt and OneSearch, use different encryption methods to keep your query private)

*** Language learning tools**

Please, select as many as apply.

13

Figure 19: Full survey as published (page 13/18)

☐ Computer-assisted language learning tools (e.g. Duolingo, FluentU, SKELL)
☐ Web-based thesaurus tools (help users to find synonyms of words e.g. thesaurus.com)
☐ Intelligent systems to aid and assess reading comprehension (e.g. Whooo's Reading, Storia)
☐ Web-based translation search engines (e.g. Linguee)

* If "other", please specify.

Do you perceive gaps in technological support for the minority/regional/lesser-used language(s) you work with?

☐ Yes
☐ No

What are the main problems (or poor results) you observe when using these language tools?

Please, select as many as apply.

☐ Gaps in the amount and variety of available applications
☐ Gaps in the quality of the tool/application (delays in responding, difficulties with special characters, language-related errors in the output etc.)
☐ Gaps in the variety of linguistic phenomena/text types covered
☐ Gaps in adaptability to systems (e.g. adaptability to iOS system)
☐ The tool and all the resources (help pages etc.) only being available in English
☐ Unavailability of resources in the language because of restrictive licences
☐ Not sure
☐ Other

* If "other", please specify.

In your opinion, what is going well when using these language tools?

In general terms, how do you evaluate the performance of the language technology tools for the minority/regional/lesser-used language(s) you work with? Please evaluate based on a four-point scale.

Please, select as many tools as apply. If you cannot evaluate for any reason, please select not applicable (N/A).

14

Figure 20: Full survey as published (page 14/18)

	1. Very poor	2. Poor	3. Good	4. Excellent	5. N/A
Proofing tools (e.g. Spell checkers, Autocorrect)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Translation tools (e.g. Google Translate)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speech recognition tools (e.g. Siri, Alexa)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parsing (e.g. PoS taggers)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search tools (e.g. Google search)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sentiment analysis and opinion analysis tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text summarization (e.g. Quillbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text mining (e.g. IBM Watson)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Language learning (e.g. Duolingo, thesaurus, bilingual dictionaries)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

If "other", please specify.

Please, choose the option that best describes the level of language technology support for the minority/regional/lesser-used language(s) you or your organisation work with.

Please, select as many tools as apply. If you do not know one or more tools, select not applicable (N/A).

	1. Very poor	2. Poor	3. Good	4. Excellent	5. N/A
Proofing tools (e.g. Spell checkers, Autocorrect)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Translation tools (e.g. Google Translate)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speech recognition tools (e.g. Siri, Alexa)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parsing (e.g. PoS taggers)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search tools (e.g. Google search)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sentiment analysis and opinion analysis tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text summarization (e.g. Quillbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text mining (e.g. IBM Watson)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

15

Figure 21: Full survey as published (page 15/18)

Language learning (e.g. Duolingo, thesaurus, bilingual dictionaries)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

* If "other", please specify.

Please indicate based on a five-point scale how frequently you use the language technology tools /applications listed below for the minority/regional/lesser-used languages you work with.

Please, select as many tools as apply.

	1. Never	2. Rarely	3. Sometimes	4. Frequently	5. Every day
Proofing tools (e.g. Spell checkers, Autocorrect)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Translation tools (e.g. Google Translate)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Speech recognition tools (e.g. Siri, Alexa)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Parsing (e.g. PoS taggers)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Search tools (e.g. Google search)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sentiment analysis and opinion analysis tools	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text summarization (e.g. Quillbot)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Text mining (e.g. IBM Watson)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Language learning (e.g. Duolingo, thesaurus, bilingual dictionaries)	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Other	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

* If "other" tool, please specify.

Predictions and visions for future

* In your opinion, what provision of resources would increase the use of language tools for the specific languages you or your organisation use?

16

Figure 22: Full survey as published (page 16/18)

Please, select as many as apply.

- ☐ A wider range of language tools for the languages I work with
- ☐ Higher-quality tools for the languages I work with
- ☐ More training of personnel dealing with such tools
- ☐ More resources (time, financial) available to work with the technology
- ☐ More outreach activities and activation outside the community
- ☐ Other

* If "other", please specify.

Which tools or applications that substantially use language technology do you want to see in the community you represent that are not available today? (we welcome any suggestion, even ideas that are not possible with current technology)?

Please indicate the best option that describes your vision for the future of languages technology.

	1. Strongly disagree	2. Disagree	3. Undecided	4. Agree	5. Strongly Agree
* In the next 10 years, there will be higher-quality language tools that deal with all the languages that concern me, including minority languages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
* In the next 10 years, there will be a wider range of language tools for European Languages	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
* In the next 10 years, language technology tools will help prevent the loss of linguistic diversity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

*** In your opinion, what would be the most relevant benefits of improving technologies for the languages you, your community or your organisation work with (including minority/regional/less-used languages)?**

Please, select as many as apply.

- ☐ Increase individuals' exposure to these languages
- ☐ Prevent minority/regional languages from disappearing
- ☐ Increase the number of speakers of those languages, including minority/regional languages

17

Figure 23: Full survey as published (page 17/18)

☐ Improve communication between native speakers
☐ Improve literacy for minority/regional languages
☐ Enhance the communication capabilities of people with disabilities
☐ Increase engagement with social, leisure and work activities in their own languages
☐ Improve online trade in countries where those languages are spoken
☐ Improve offline trade (i.e. not e-commerce) in countries where those languages are spoken
☐ Other

* If "other", please specify.

If you have any comments/suggestions, please let us know.

* Can we contact you to arrange a possible follow-up discussion?

☐ Yes
☐ No

* What is your e-mail address?

What is your name?

☐ By clicking on 'Submit', I agree that my personal data (email address and/or name) can be used according to the Privacy Policy of the European Language Equality (ELE) project.
[ELE Privacy Policy.pdf](#)

18

Figure 24: Full survey as published (page 18/18)

B. Tables for Analysis

Languages	Number of articles
English	6,408,063
Swedish	2,887,160
German	2,631,755
French	2,373,765
Dutch	2,071,178
Spanish	1,730,433
Italian	1,725,74
Polish	1,496,337
Portuguese	1,077,16
Catalan	689551
Serbian	651405
Norwegian (Bokmål)	569178
Finnish	519488
Hungarian	494290
Czech	491751
Turkish	448125
Romanian	424465
Basque	382477
Tatar	320512
Bulgarian	276529
Danish	270625
Slovak	237989
Estonian	223118
Croatian	209449
Lithuanian	200402
Greek	199303
Galician	177585
Slovene	174108
Norwegian (Nynorsk)	160302
Welsh	133859
Asturian	128391
Macedonian	118115
Latvian	109589
Bosnian	88176
Occitan	86976
Albanian	84234
Breton	70279
Venetian	68734
Piedmontese	65872
Luxembourgish	60302
Irish	55791
Icelandic	53466
Lombard	49621
West Frisian	47002
Aragonese	40667
Silician	26189
Scottish Gaelic	15540

Yiddish	15298
North Frisian	14450
Upper Sorbian	13761
Faroese	13637
Emilian-Romagnol	12943
Ligurian	10810
Northern Sami	7778
Sardinian	7121
Võro	5936
Kashubian	5398
Picard	5262
Franco-Provençal	5174
Manx	5070
Cornish	4960
Maltese	4307
Saterland Frisian	4049
Mirandese	3865
Inari Sami	3724
Livvi-Karelian	3698
Ladino	3606
Friulian	3460
Lower Sorbian	3303
Aromanian	1267
Latgalian	1013
Romani	706

Table 2: Wikipedia language versions for the ELE languages, retrieved from <https://wikistats.wmcloud.org/display.php?t=wpNovember2021>

ELE Languages	Number of recordings in Lingua Libre
French	231235
Polish	54846
Romanian	19400
English	19353
German	14500
Occitan	14057
Swedish	7735
Languedocien	5327
Portugues	5266
Gascon	4887
Spanish	4674
Italian	3529
Basque	3276
Catalan	2265
Macedonian	1959
Dutch	1330
Finnish	1174
Welsh	748
Breton	693

British English	202
Luxembourgish	86
Aromanian	34
Lemosin	32
Norwegian	29
Greek	18
Upper Saxon German	12
Czech	10

Table 3: Number of recordings on Ligua Libre for languages of the ELE consortium, retrieved from <https://lingualibre.org/wiki/LinguaLibre:Stats/Languages> on November 9 2021

ELE Languages	Number of Lexemes in Wikidata
Russian	101322
Estonian	83208
English	71302
Swedish	35600
German	25574
Basque	22913
Slovak	16475
Czech	13106
Bokmål	12552
French	12243
Danish	12137
Spanish	5744
Portuguese	3105
Nynorsk	2850
Polish	2550
Luxembourgish	852
Italian	683
Finnish	623
Breton	283
Maltese	212
Northern Sami	199
Dutch	175
Bulgarian	161
Welsh	137
Catalan	122
Yiddish	121
British Sign Language	119
Croatian	117
Latvian	108
Hungarian	85
Occitan	79
Romanian	59
Albanian	54
Modern Greek	49
Faroese	41
Galician	31
Serbian	28

Kashubian	26
Friulian	23
Irish	21
Cornish	19
Manx	17
Nepali	17
Middle Danish	17
Norman	16
Aromanian	15
Sicilian	14
Icelandic	13
Venetian	13
Crimean Tatar	12
West Frisian	11
Lithuanian	10
Sardinian	10

Table 4: Number of Lexemes in Wikidata for the languages of the ELE consortium, retrieved from <https://ordia.toolforge.org/language> on December 1 2021

Wikimedia Belgium
 Wikimedia Česká republika
 Wikimedia Danmark
 Wikimedia Deutschland
 Wikimedia Eesti
 Wikimedia España
 Wikimedia Suomi
 Wikimédia France
 Wikimedia Italia
 Wikimedia Nederland
 Wikimedia Austria
 Wikimedia Polska
 Wikimedia Portugal
 Wikimedia Sverige
 Wikimédia Magyarország
 Wikimedia Community User Group Albania
 Basque Wikimedians User Group
 Wikimedia Community User Group CEE Spring
 GLAM Macedonia User Group
 Wikimedia Community User Group Greece
 Wikimedia Community Ireland User Group
 Wikimedia Community of Kazakh language User Group
 Wikimedians of Latvia User Group
 Wikimedia Community User Group Malta
 Wikimedians of Romania and Moldova User Group
 Wikimedians of Slovakia User Group
 Wikimedians of Albanian Language User Group
 Wikipedians of Slovenia User Group
 Wikimedia Small Projects in Spanish User Group
 Wikimedia Community User Group Wales

Wikimedians of Republic of Srpska
In addition, there are further chapters located in Europe:
 Wikimedia CH
 Wikimedia UK
 Wikimedia Serbia
 Wikimedia Ukraine
 Wikimedia Norge

Table 5: List of Wikimedia chapters and user groups in the EU and dealing with European languages

C. Additional tables and graphs

Countries	Answers count	%
Spain	4	18,2
Germany	3	13,6
France	3	13,6
Hungary	2	9,1
Russia	2	9,1
Slovak	1	4,5
Ireland	1	4,5
Denmark	1	4,5
Bulgaria	1	4,5
Malta	1	4,5
Macedonia	1	4,5
US	1	4,5
Wales	1	4,5

Table 6: Breakdown of answers count to the question “In which country are you based?”

Types of organisations	Answers count	%
Education/research	11	50
N/A	6	27,3
Advocacy and education	1	4,5
Basque Wikimedians User Group	1	4,5
Large Enterprise	1	4,5
Professional associations	1	4,5
SME	1	4,5

Table 7: Breakdown of answers count to the question “Which of the following best describes the type of organisation you work for?” (Example of mandatory single choice question)

Organisations
EWKE – BWUG
Wikimedians of Slovakia/Slovak Wikipedia
German wikipedia user
French Wikipedia, Wikimedia Community of Saint Petersburg User Group
Wiktionary
Catalan Wikipedia
Macedonian Wikipedia
Wikimedia search developers
Aragonese Wikipedia
Wikidata
Wikimedians on the island of Ireland
Hungarian Wikipedia and Wikidata.
North Frisian Wikipedia
Wikipedia, Wiktionary
Wikidata, Danish Wikipedia
Bulgarian Wikipedia
French Wiktionary
Wales User Group
cywiki (Welsh / Cymraeg Wikipedia)
Lingua Libre
Hungarian Wikipedia
Wikidata, Maltese Wikipedia

Table 8: Breakdown of answers to the question “Which community are you representing? (e. g., Wikidata, Italian Wikipedia, User Groups etc)?”

Languages	Answers count	%
English	14	63,6
French	7	31,8
Spanish	6	27,3
Danish	5	22,7
German	5	22,7
Hungarian	5	22,7
Dutch	4	18,2
Irish	4	18,2
Italian	4	18,2
Norwegian	4	18,2
Slovak	4	18,2
Swedish	4	18,2
Bulgarian	3	13,6
Croatian	3	13,6
Czech	3	13,6
Estonian	3	13,6
Finnish	3	13,6
Greek	3	13,6
Latvian	3	13,6
Lithuanian	3	13,6
Polish	3	13,6
Portuguese	3	13,6
Romanian	3	13,6
Icelandic	2	9,1
Maltese	2	9,1
Russian	2	9,1
Slovenian	2	9,1
Basque	1	4,5
Catalan	1	4,5
Chinese	1	4,5
Macedonian	1	4,5
Mandarin	1	4,5
Welsh	1	4,5

Table 9: Breakdown of answers to the question “For which language (s) you, your community or your organisation use language technology tools (e.g., Translation tools, Spell/grammar checkers, web search engines, social media, language learning tools)?”

Language Technologies	Answers counts	%
Parsing tools		
Dependency or constituency parsing systems	1	4,5
Part-of-speech taggers of any type	3	13,6
Proofing tools		
Autocorrect tools	4	18,2
Grammar checkers	5	22,7
Spell checkers	6	27,3
Search tools		
Cross-language search engines	0	0,0
Customer-build search engines	1	4,5
Domain-specific search engines		
Generic search systems freely on the web	3	13,6
Multilingual search engines	3	13,6
Multimedia search engines	0	0,0
Ontology tools	1	4,5
Web-based question-answering systems	3	13,6
Speech technologies		
Text-to-speech systems	3	13,6
Voice user interfaces	3	13,6
Translation tools		
Computer-assisted translation tools	6	27,3
Custom-built translation engines	1	4,5
Generic translation tools freely available on the web	8	36,4
Terminology management applications	1	4,5

Table 10: Breakdown of answers count to the question “Which tools/applications do you use with these languages?”

Language Technologies	Answers counts	%
Proofing tools		
Autocorrect	1	12,5
Grammar checkers	2	25
Spell checkers	3	37,5
Search tools		
Customer-build search engines	1	12,5
Generic search systems freely on the web	2	25
Multilingual search engines	4	50
Web-based question-answering systems	2	25
Translation tools		
Computer-assisted translation tools	1	12,5
Custom-built translation engines	3	37,5
Generic translation tools freely available on the web	3	37,5
Other tools		
Lingua Libre	1	12,5

Table 11: Breakdown of answers to the question: “Which tools/applications do you use with these minority/regional/lesser-used languages? if “other”, please specify.”

For Google, the webmail's outer interface is not available in Basque, referring back to Spanish, even when not required.
Search engines detecting lexemes, instead of words
More time would be best, for sure. More learning by people rather than AI.
linguaLibre, living dictionaries
Better voice recognition
Stemmers are very helpful in improving search results for more highly inflected languages.
Online date apps Preferably AGDA or at least ISO/SQL.
North Frisian is not yet represented in Unicode, although it was requested several years ago.
North Frisian is not yet represented in Google translate.
Multi-translation tools: capacity to translate articles to and from multiple languages
Promote communities of learning languages in europe through technology
I would love to have a tool in which you can write in your own language and that would display (automatically) translated what you wrote in the mother tongue of the reader, so that a text can be written by several people speaking different languages and read by other people reading other languages.
high-quality open-source translation tools, better support for agglutinative languages in various tools (e. g., in Android typing assistants), sentiment analysis tools for Hungarian, better grammar and style checkers for Hungarian, translation memory / translation workflow support
Spell checker and speech to text.

Table 12: Full list of answers to “Which tools or applications that substantially use language technology do you want to see in the community you represent that are not available today? (we welcome any suggestion, even ideas that are not possible with current technology)?”