The Role of National Language Institutions in the Digital Age

EFNIL

European Federation of National Institutions for Language

Željko Jozić / Sabine Kirchmeier (eds.)

The Role of National Language Institutions in the Digital Age

Contributions to the EFNIL Conference 2021 in Cavtat

ISBN 978-963-9074-92-7 (Print)
978-963-9074-93-4 (PDF)

Published by
Nyelvtudományi Kutatóközpont / Hungarian Research Centre for Linguistics
Benczur u. 33
H-1068 Budapest

Layout and typesetting: Joachim Hohwieler
Cover design: Norbert Cußler-Volz

Preface

This volume contains the talks given at EFNIL's 18th annual conference that took place in Cavtat/Croatia on 6th–8th October 2021. The conference was a cooperation between the Institute for Croatian Language and Linguistics and the European Federation of National Institutions for Language (EFNIL).

The theme of the conference was:

The Role of National Language Institutions in the Digital Age.

Many EFNIL institutions not only cater for the national language but also care for minority languages, regional languages, and sign languages in their country. In many cases, speakers of less frequently spoken languages do not have access to the digital services that are available to users of more widespread languages. This is a problem not only for their participation in the democratic debate and activities in their countries but also for their use of digital public services and efficient use of digital tools.

There is not a one-to-one relationship between frequently spoken languages and national languages in terms of the number of speakers. There are regional languages, Catalan for instance, that have more speakers than national languages such as Lithuanian or Icelandic, and therefore national languages, regional languages, and minority languages are facing the same challenges and need to work together.

The welcome addresses by Radovan Fuchs, Minister of Science and Education, and Nina Obuljen Koržinek, the Minister of Culture and Media, of the Republic of Croatia, both stress the importance of language for the cultural and national identity of individual countries and the cultural diversity of Europe, thereby underpinning the necessity to strengthen the digital presence of all languages on equal terms.

Fortunately, the European Commission is very well aware of the situation for less frequently spoken languages. Head of Sector Multilingualism at DG/CONNECT of the European Commission, Philippe Gelin, presented several projects such as ELRC (European Language Resource Coordination) and ELE (European Language Equality) aiming at paving the way for the inclusion of more and more languages in the digital world.

The first article in this volume describes the ELE (European Language Equality) project and its goal to achieve digital language equality for all European languages by 2030. The next two articles – on the situation for Greenlandic and the Sami languages – illustrate how difficult it is for languages with only 50,000 and 20,000 speakers, respectively, to overcome the commercially motivated barriers created by multinational companies such as Microsoft and Google. Providing services like spell checkers or machine translation for these languages does not seem to represent a sufficiently interesting business case, and therefore they are simply ignored.

The following articles describe initiatives taken in Germany, Poland, Romania, Slovenia, and Switzerland as well as the Netherlands and Belgium with regard to language resources and language technology to support language communities and ensure their prevalence in the digital sphere. They all stress the role that national institutions for language play in this context.

Finally, this volume contains a report based on a survey regarding the situation of European languages in the public space (ELIPS). The survey, which is the first of its kind in Europe, was conducted by a project group composed of EFNIL members with the aim of collecting information about how public institutions communicate with their citizens. It features an extensive analysis of the initiatives that are taken in various countries in terms of legislation, tools, methods, and best practices to improve public communication, and represents a huge pool of inspiration for everyone who is interested in creating good and clear information for all citizens. This aspect is also relevant for digital communication and language technology as experience shows that clear language and consistent terminology have a direct impact on the quality of digital tools.

We believe that this volume represents important knowledge about the digital language situation in Europe and we thank all speakers and contributors to this volume. Finally, we are deeply grateful to the Institute of Croatian Language and Linguistics for introducing the topic and hosting the conference.

Željko Jozić Sabine Kirchmeier

Contents

EFNIL project report

Appendix

Johan Van Hoorde

Introduction

Dear representatives of the European Commission,
Dear special guests and invited speakers,
Dear colleagues,

It is a pleasure and privilege to stand here and welcome you to the 18th conference of the *European Federation of National Institutions for Language* EFNIL, the collaborative platform of the official languages of various European nations.

At our previous conference two years ago in Tallinn I admitted that it was with some nervousness that I welcomed you. This is certainly also the case today, be it for different reasons. Due to the covid pandemic we have not been able to organise a conference for two years now. And finally here we are. A new live conference! What a pleasure to stand in front of you, see your faces, hear your voices, feel your enthusiasm!

For our Croatian hosts at the *Institute of Croatian Language and Linguistics*, preparing for this event has been particularly difficult, due to the uncertainties surrounding travel and other problems related to the pandemic. They deserve a warm round of applause for all of their efforts, with this marvellous result.

Given this uncertainty, we had to keep all possibilities open and opted for a hybrid conference with both live and online attendance. Please allow me to also welcome all our colleagues and friends who are with us online. I hope they will feel part of our language family and feel something of the friendly, collaborative atmosphere.

And indeed, ladies and gentlemen, here we are in the beautiful Dalmatian area of Croatia. This evening we will have the opportunity to visit the old historic centre of Dubrovnik. There is a link between Dubrovnik and Tallinn, where our last conference took place. Both cities are unique historical places, recognised by UNESCO as World Heritage Sites.

Dubrovnik has always been a maritime centre of commerce and as such, a place of encounter between people of different origins, with a variety of languages as well as different cultural and religious traditions. This makes this area a symbolic place for an EFNIL conference. EFNIL tries to be a model of understanding and collaboration between language communities within Europe.

It is with great pleasure that I welcome the representatives of the European Commission. And indeed, it is with great pleasure that we heard the video messages by the Minister of Culture and Media and by the Minister of Science and Education of the Republic of Croatia, Ms Obuljen and Mr Fuchs respectively. We consider the presence and contributions of all of these authorities as an honour and thank them for their lively interest in our work.

I would have loved to have welcomed Ms Obuljen and Mr Fuchs in person, but we understand that their work agendas did not allow them to come to Cavtat. I thank them for their nice words.

Draga ministrice Obuljen Koržinek, dragi ministre Fuchs, hvala vam odsrca na ohrabrujućim riječima koje ste nam uputili kao predstavnici Vlade Republike Hrvatske. Drago nam je i ponosni smo što imamo Republiku Hrvatsku, a osobito hrvatski jezik kao dio naše europske jezične obitelji.

Let me turn back to English now and to the topic of this conference. Today and tomorrow we will be discussing the role of our national language institutions in the digital age. The relationship between the mission of our institutions and digitalisation is a complex one. It is obvious that the digital revolution has had and is having a great impact on research and the scientific study of language, as well as on the production of language resources such as dictionaries, terminology databases, and text corpora. Our work is clearly influenced by digitalisation, and the nature of this influence on language planning and research will be one of the subthemes during this conference.

But we also have to look at the other side, that is the ICT sector and the challenges it has to cope with. Among these there is certainly also the language challenge, if solutions for communicative and other needs are supposed to be universally available for all consumers worldwide. That means that language resources and linguistic expertise can help the industry to improve its products, to cross language barriers and to increase the power and impact of its innovative solutions. That is a second perspective that will be discussed today and tomorrow.

The third perspective is political in nature in relation to language use. How is the digital revolution influencing the status and position of our languages? Almost all products, solutions and technologies are available in English and in some other big and powerful languages, but not, or not automatically, in languages with a smaller home market. This certainly means that the digital revolution risks creating or reinforcing power differences between languages and language communities. For relatively small languages there is the risk of loss of functional domains if people are offered solutions and innovative types of support in English but not or not to the same extent and with the same quality and impact in their home language. For these languages it is not evident that the free market will develop all solutions itself, which as a consequence could imply a more active role for the public sector. The issue of equal opportunities for all languages is, without a doubt, one of the bigger political challenges and will,, of course, be a topic of discussion at this conference.

Above I described the socio-political problem from the perspective of the interest of our languages and their future. The same problem can also be approached from another perspective, one that is even more important, at least from a social point of view. It will be clear to everyone that technological innovation not only risks creating power and functionality differences between languages but certainly also between individuals, between a social and knowledgeable upper class and less favoured social classes. The covid-19 pandemic has made this very clear. Schools were closed down and live lessons in the classrooms had to be replaced by online lessons with modern communication platforms. Many pupils from lower social classes found themselves excluded because they had no laptops or good internet connections. This social aspect should be taken into account if our policies are supposed to be committed to inclusive citizenship, avoiding all types of exclusion and social discrimination. Without a doubt, language is one of the dimensions – albeit not the only one – of this social challenge. Our national languages are the language varieties that are by far the most widespread among the population at large and, hence, guarantee the best possible access to information and knowledge. To put it simply: if products and services were available only in English and some other privileged languages but not in all the others, this would reinforce the social gap between an elite that can use them and all the others that lack the language competence and would remain deprived of them.

Dear participants, there will be ample space to discuss all of these important aspects from all of these different angles. There is one political aspect that is not explicitly part of the programme but to which I want to draw your attention, be it only briefly. In discussing language planning in the digital age, we tend to focus almost exclusively on the influence these two phenomena can have on each other, that is what technology can mean for languages and language research and how language expertise and language resources can support digital innovation. There is a danger, that in doing so, we lose track of what could perhaps be the most powerful impact.

This regards the way the digital age tends to modify the social habits and interaction patterns of human beings and, in doing so, their language needs and indeed the very construction of their cultural and social identities. I am referring to the nexus between offline and online communication and human interaction. Up to fifty years ago or so, all our social and communicative interaction took place in a physical space, composed of concrete geographical places, from local to regional to national and – only for a very small minority – beyond the boundaries of the nation state. This is no longer the case, as the famous Spanish sociologist Manuel Castells has demonstrated. The *space of places* which has been the scene for human interaction is being replaced by what he calls a ***space of flows***, where offline and online interaction intertwine. Modern humans are part of complex and variegated networks of both offline and online interaction. As a result our social and personal identities are increasingly fragmented, idiosyncratic identities that are, to a much lesser extent, defined or inspired by our direct social environment, i.e. our home town, our language community, our nation state. In other words, the new social interaction patterns challenge the old paradigm of largely monolithic identities, with one language, one set of values and norms, and one model of what is socially acceptable, shared by almost all members of a geographically based community.

Needless to say, this new social reality will have serious consequences for the linguistic competence needed by individuals to be fully part of this globalising digital age. More and more they will need and use complex linguistic repertoires that go beyond one (standard) variety of one national language. This change might be the most powerful one that will force us as language institutes to reconsider the status and position of our languages in view of the social and communicative needs of our citizens.

Dear participants, I hope to have convinced you of the importance of the topic of this conference as well as of the richness of the aspects and perspectives that are potentially involved. Let's now start with the conference. We have an exciting programme with stimulating keynote speakers and excellent national reports.

I am delighted to introduce our first keynote speaker, Mr Philippe Gelin. He is head of the sector on Multilingualism at the DG for Communication Networks, Content and Technology, better known as DG Connect. We consider his presence with us today as a sign of support from the European Commission and as a promising base for further contacts and collaboration. Dear Mr Gelin, you are, without a doubt, an expert in the field of digital technologies and how these influence communicative and social behaviour. You have long-term experience in all relevant subdomains, be it scientific research, applied industrial research or technology development and – for many years – as a policy agent for the European Commission. Mr Gelin, the floor is yours.

References

Castells, M. (2020): Space of flows, space of places: Materials for a theory of urbanism in the information age. In: LeGates, R. T./Stout, F. (eds): *The city reader.* London: Routledge, 229-240.

Radovan Fuchs

Greetings from the Minister of Science and Education of the Republic of Croatia

Dear President of EFNIL, dear directors of the European language institutions, dear participants, it is a great pleasure to welcome you all to the opening of the 18th EFNIL Conference jointly organized by the Institute for the Croatian Language and Linguistics and the European Federationof National Institutions for Language.

The Ministry of Science and Education greatly values and supports the work being done by the Institute for the Croatian Language and Linguistics because we are well aware of the importance of the language for our cultural and national identity. The role of EFNIL is essential in supporting research into official European languages; the promotion of linguistic and cultural diversity within the European Union is also the main mission of your organization, but this mission is shared and should be fostered and promoted by all of us. It is especially important to be aware of this event in this technological era and I am glad that in the coming days of the conference you will be exploring challenges and opportunities that this digital era presents for language research.

Teaching the national language at all educational levels in schools in order to promote written and oral competence is necessary to enable people to play a full role in society. Language is the medium by which we can directly participate in all forms of communication, exchange of cultural elements, knowledge, beliefs, art, morals, rights, and habits that make culture distinctive.

I believe that this important topic will benefit from this conference, both in terms of broadening theoretical insights and advancing the state of the art in language research and resources, and in finding new ways to preserve and promote linguistic and cultural diversity within the European Union.

Let this event be an inspiration and incentive for the exchange of new knowledge and experience. With this in mind I want to thank all the expert participants at this event once again, and I wish you all some very successful following days in the beautiful Dubrovnik area. Thank you very much for your attention.

Nina Obuljen Koržinek

Welcome address of the Minister of Culture and Media of the Republic of Croatia

Dear President of the European Federation of National Institutions for Languages, dear directors, dear participants!

First of all, I would like to say that I am very sorry that I was not able to meet you in person and greet you in person in Cavtat for this very important conference, the 18th Conference of the European Language Institutions. We all know how important language is, not only as a means of communicating values and beliefs but also as a sign of identity and as a vehicle for conveying our culture and for communicating different cultures. In the process of preserving languages, the role of national institutes has a very important place and I would like to say that for us in the field of culture, the work of national institutes and everything you do to preserve and develop cultures is extremely important.

Your topic for this conference is the position and role of language in the digital age, which we know is particularly relevant and important, and I am convinced that the deliberations that you will have in the coming two days will bring you closer when sharing your experiences and findings, both scientific and other, on what can be done so that we support the study of languages and the role of languages in the digital age.

Once again, I am sorry that I was not able to join you, but I hope that you will have very productive discussions and that you will enjoy your time in Cavtat.

Georg Rehm/Federico Gaspari/German Rigau/Maria Giagkou/
Stelios Piperidis/Annika Grützner-Zahn/Natalia Resende/
Jan Hajic/Andy Way

The European Language Equality Project: Enabling digital language equality for all European languages by 2030

Abstract

The EU project European Language Equality is currently preparing a strategic research, innovation and deployment agenda and roadmap which will provide a detailed plan and strategic recommendations on how to achieve digital language equality in Europe by 2030. This article presents an overview of the project, our definition of digital language equality and preliminary results using the associated DLE metric. The final project documentation including the strategic agenda will be handed over to representatives of the European Union in mid-2022.

1. Introduction: Natural Language Processing and Language Technology in Europe

Language Technology (LT) is one of the most important AI application areas with a fast-growing economic impact. Current LT (NLP, Speech, Multimodal, etc.) supports many advanced applications which would have been unthinkable only a few years ago. In fact, the LT community in multiple sectors (Machine Translation, Text Analytics, Speech, Language Resources, etc.) is developing new powerful deep learning techniques, tools and large multilingual pre-trained language models that will revolutionize many language-related tasks and support improved ways of communication, including across languages. Even just five years ago, only a few firm advocates would have predicted the recent breakthroughs that have resulted in systems that can translate without parallel corpora (Artetxe et al. 2019), create image captions (Hossain et al. 2019), generate full text claimed to be almost indistinguishable from human prose (Brown et al. 2020), generate theatre play scripts (Rosa et al. 2020) and create pictures from textual descriptions (Ramesh et al. 2021) as well as systems able to deal with unseen tasks (Min et al. 2021; Sanh et al. 2021; Wei et al. 2021; Ye et al. 2021). While forecasting the future of LT and language-centric AI is a challenge, it is, we believe, safe to predict that even greater advances will be achieved in all LT research areas and domains in the near future.

However, despite claims of ‘human parity’ in many LT tasks (e.g. in Machine Translation, by Wu et al. 2016 and Hassan et al. 2018), Deep Natural Language Understanding (NLU) is still an open research problem which is far from being solved since all current approaches have severe limitations (Bender et al. 2021). Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pre-trained language models and self-supervised systems opens up the way to leverage LT for less digitally supported languages. For the first time, a single multilingual model has outperformed the best specially trained bilingual models on news translations, i.e. one multilingual model provided the best translations for both low- and high-resource languages, showing that the multilingual approach is indeed the future of MT (Tran et al. 2021), especially if high-quality MT is really going to be rolled out for all of the world’s 7000+ languages. Indeed, some believe this to be achievable in relatively short periods of time; Meta CEO Mark Zuckerberg recently asserted “the ability to communicate with anyone in any language: that’s a superpower people have dreamed of forever, and AI is going to deliver that within our lifetimes” (cf. the accompanying blog by Edunov et al. 2020). For that to be achievable, the development of these new LT systems would not be possible without sufficient resources (experts, data, computing facilities, etc.) as well as the creation of carefully designed and constructed evaluation benchmarks and annotated datasets for every language and domain of application.

Unfortunately, there is no equality in terms of tool, resource and application availability across languages and domains. Although LT has the potential to overcome the linguistic divide in the digital sphere, most languages are neglected for various reasons, including an absence of institutional engagement from decision makers and policy stakeholders, limited commercial interest or insufficient resources. For instance, Joshi et al. (2021) and Blasi et al. (2021) have recently looked at the relation between the types of languages, resources and their representation at NLP conferences over time. Disappointingly, but perhaps not altogether unexpectedly, only a very small number of the 7000+ languages of the world are represented in the rapidly evolving LT field. A growing concern is that due to unequal access to digital resources – especially as larger and larger AI models are advocated as the way forward – only a small group of big technology companies (mostly non-European) and elite universities will lead modern LT development (Ahmed/Wahed 2020). More alarming still is the report by Bromham et al. (2021), who found that 37% of the world’s 6,511 languages which they investigated (i.e. approximately 90% of the total number of languages in the world) are considered to be threatened or endangered (i.e. losing first-language speakers or only spoken by adults, without child learners), while 13% were placed in the even less enviable category of “sleeping” (i.e. no longer spoken as first languages). Overall, this means that around 50% of the investigated languages (i.e. over 3,000 of them across the world) face serious risks of extinction, potentially within a generation, if not imminently.

To unleash the full potential of LT and ensure that no users of these technologies are disadvantaged in the digital sphere simply due to the language they speak, we argue that there is a pressing need to facilitate long-term progress towards multilingual, efficient, accurate, explainable, ethical, fair and unbiased language understanding and communication. In short, we must ensure transparent Digital Language Equality (DLE) in all areas of society, from government to business to citizens. In the 21st century, language cannot be an impediment to accessing information, and LT is the only feasible way to overcome language barriers while preserving the rich cultural diversity and linguistic rights held dear by all European citizens.

The remainder of this paper is organized as follows. Section 2 describes the setup and goals of the EU project European Language Equality, the first results of which are reported on in this article. Section 3 explains the methodology applied in the project. Section 4 describes our results to date, and Section 5 concludes the paper, providing the expected next steps in the ELE project and beyond.

2. European Language Equality (ELE): Context and goals

In a plenary meeting on 11th September 2018, the European Parliament adopted, with an overwhelming majority, a joint ITRE/CULT report, “Language equality in the digital age”, with a resolution that included over 40 recommendations. These concern the improvement of the institutional framework for LT policies at EU level, EU research policies, education policies to improve the future of LTs in Europe, and the extension of the benefits of LTs for both private companies and public bodies (European Parliament 2018). In particular, the resolution highlighted many important areas, e.g. it called on the Commission “to establish a large-scale, long-term coordinated funding programme for research, development and innovation in the field of language technologies, at European, national and regional levels, tailored specifically to Europe’s needs and demands”. While the European Commission has been funding LT for many years now, it is the case that LT has not really been at the centre of European policy making, and the ITRE/CULT report says that it should be.

While the 24 official EU languages have been granted equal status politically, technologically they are far from equally supported; in addition, there are several regional and minority languages that have traditionally suffered from limited support, especially to future-proof their use and very existence in the digital age. The goal of the €1.8 million EU-funded project European Language Equality (ELE)[1] is the systematic and inclusive development of an all-encompassing strategic research, innovation and implementation agenda (SRIIA) and roadmap for achieving full DLE in Europe by 2030, exactly as recommended in the ITRE/CULT report.

[1] https://european-language-equality.eu/.

3. Methodology

Developing a strategic research, innovation and implementation agenda and roadmap for achieving full DLE in Europe by 2030 involves many stakeholders with different perspectives. Accordingly, the ELE project – led by DCU, and with DFKI, Charles University, ILSP and EHU/UPV as core members – has put together a large consortium of 52 partners who, together with the wider European LT community, are preparing the different parts of the strategic agenda and roadmap.

On a general level, we distinguish between input for the agenda and roadmap generated by the consortium, and input generated by organizations not participating as partners in the project. The results and feedback gathered internally from consortium partners as well as from external stakeholders were systematically collected and being analysed prior to its eventual inclusion in the research agenda and roadmap (SRIIA), a coherent and convincing strategy which was delivered to the Commission in June 2022 demonstrating how DLE can be achieved for all European languages by 2030.

All work strands in the project produce input for the strategic agenda. We are concentrateing on two distinct aspects: (i) collecting the current state of play (2021/2022) of LT support for the more than 70 languages under investigation, largely by the 32 National Competence Centres in our sister project, the European Language Grid (ELG);[2] and (ii) strategic and technological forecasting, i.e. estimating and envisioning the future situation in 2030 and beyond. Furthermore, we distinguish between two main stakeholder groups: LT developers (industry and research) and LT users as well as consumers. Both groups are represented in ELE by several networks (e.g. EFNIL, ELEN, ECSPM) and associations (e.g. ELDA, LIBER), who produced one report each, highlighting their own individual needs, wishes and demands towards DLE. The project's industry partners produced four "deep dives" with the needs, wishes and visions of the European LT industry regarding Machine Translation, Speech, Text Analytics and Data, all available on the project website. We also organized a larger number of surveys (inspired by Rehm/Hegele 2018) and consultations with stakeholders who are not represented in the consortium.

Our methodology is, thus, based on a number of stakeholder-specific surveys as well as collaborative document preparation that also involves technology forecasting. Both approaches are complemented by the collection of additional input and feedback through various online channels. The two main stakeholder groups (LT developers and LT users/consumers) differ in one substantial way: while the group of commercial or academic LT developers is, in a certain way, closed and well represented through relevant organizations, networks and initiatives in our

[2] https://www.european-language-grid.eu/.

consortium, the group of LT users is an open set of stakeholders that is only partially represented in our consortium. Both stakeholder groups have been addressed with targeted and stakeholder-specific surveys.

3.1 Digital Language Equality

Based on various exchanges with a range of external stakeholders, a preliminary working definition of DLE was formulated to further drive our activities:

> Digital Language Equality is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age.

The definition is further based on a set of modular quantifiers that reflect the level of support of LTs for all European languages as an essential requirement to achieve full DLE in Europe by 2030. The preparation of a strategic plan to achieve this requires the accurate and up-to-date description of the current state of technology support for Europe's languages, also to facilitate the identification of gaps and issues with regard to LTs. While the proposed DLE definition is firmly rooted in the state of the art, it will also serve the needs of the languages targeted in the project and the expectations of the relevant language communities in the future. The preliminary definition is modular and flexible, i.e. it consists of well-defined (separate and independent, but tightly integrated) quantifiers, measures and indicators; for reasons described in Section 3.2, the definition is also compatible with the ELG (Labropoulou et al. 2020; Rehm et al. 2020).

The DLE definition provides the basis to compute an easy-to-interpret metric for individual languages, which enables the quantification of the level of technological support for a language and, crucially, the identification of gaps and shortcomings that hamper the achievement of full DLE. This approach enables direct comparisons across languages, tracking their advancement towards the goal of DLE, and facilitates the prioritization of needs, especially to fill existing gaps.

The DLE metric (Gaspari et al. 2022; Grützner-Zahn/Rehm 2022) is defined as a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism, tracking its progress towards the goal of DLE. The metric is computed for each language on the basis of various factors, grouped into *technological factors* (technological support, e.g. available language resources, tools and technologies) and *contextual factors* (e.g. societal, economic, educational, industrial).

The first set of technological factors concern the availability of Language Resources and Technologies (LRTs), as well as the organizations and projects covering specific languages (see Appendix A.1). Following the ELG categorization and metadata schema, these technological factors are divided into six main categories: (i) tools and services, (ii) corpora, (iii) language models and computational grammars (i.e. language descriptions), (iv) lexical and conceptual resources, (v) projects and (vi) organizations.

The second set of measures consists of contextual factors, which do not refer to strictly technological, linguistic or language-related indicators but rather have to do with general conditions and situations of the broader context of the respective language communities. The identification of these contextual factors has built on a number of diverse sources and past projects, including the STOA (2017) report, the META-NET White Paper series Europe's Languages in the Digital Age (Rehm/Uszkoreit 2012),[3] EFNIL's European Language Monitor (ELM),[4] the FLaReNet report (Calzolari et al. 2011), the META-NET Strategic Agenda for Multilingual Europe 2020 (Rehm/Uszkoreit 2013) and the Digital Language Diversity Project.[5] The preliminary list of contextual factors that contribute to the computation of the DLE metric was formulated in early 2021. Appendix A.2 lists the 72 factors, clustered into 12 categories.

Note that there is evidence that an interaction of several factors (including non-linguistic ones) seems to be beneficial. For example, using three geographical and economic factors (gross domestic product (GDP), size of the language community and geographic proximity), Faisal et al. (2021) investigated the geographical representativeness of NLP datasets, with a view to discovering the extent to which NLP datasets match the expected needs of language speakers. Given that most of the data sets came from countries considered to be economically prosperous, the best predictive value was GDP, but better predictions were achieved when taking GDP and geographic proximity into account.

We have recently refined the DLE definition and the related metric, with a focus on finalizing the list of contextual factors. After considerable effort to determine reliable sources of demographic and statistical information from which the required data can be pulled to compute the DLE metric for all languages of Europe, 26 of the 72 contextual factors (see items in red in Fig. 1) were excluded due to missing data. This affected especially factors from the classes "research & development & innovation", "society" and "policy". Data about policies are mainly too broad and just represent whether policies exist or not. The class "society" included factors about diversity which are difficult to quantify. The problem of missing data in this area was already mentioned in the AI Index report (Zhang et al. 2021). The factors excluded from the class "research & development & innovation" mainly covered specific figures about the research environment of LTs, while broader figures about the research situation of the whole country independent of research areas are available.

[3] http://www.meta-net.eu/whitepapers/.

[4] http://efnil.org/projects/elm.

[5] http://www.dldp.eu.

Economy	**Education**	**Funding**	**Industry**	**Law**	**Media**	**Online**	**Policy**	**Public Administration**	**R&D&I**	**Society**	**Technology**
Size of the economy	🟥 Higher Education Institutions operating in the language	Public funding available for LT/ NLP/AI research projects	Companies developing LTs	🟥 Copyright legislation and regulations	Publicly available subtitled or dubbed media outcomes	Digital libraries	Presecne of local, regional or national strategic plans, agendas, etc.	Languages of public institutions	Innovation capacity	🟥 Importance, relevance or recognition of the language	🟥 Open source technology in LT
Size of the LT/ NLP market	🟥 Proportion of higher education conducted	Venture capital available	Start-ups per year	Legal status and legal protection	Publicly available transcribed podcasts	Impact of language barriers on e-commerce	🟥 Recognition and promotion of the LR ecosystem	Available public services in the language	Research groups in LT	Fully proficient speakers	Access to computer, smartphones, etc.
Size of the language service and translation or interpreting market	Academic positions in relevant areas	Public funding for interoperable platforms	Start-ups in LT/AI			🟥 Digital literacy	🟥 Consideration of regional or national bodies for the citation of LRs		🟥 Research groups/ companies predominantly working on the respective language	Digital skills	Digital connectivity and Internet access
Size of the IT/ICT sector	🟥 Academic programmes of study in LT					Wikipedia pages	🟥 Promotion of regional, national or international cooperation		🟥 Research & Development staff involved in LT	Size of language community	
Investment instruments into AI	Literacy level					🟥 Websites with content available exclusively in the language	🟥 Public and community support for the definition and dissemination of resource production best practices		🟥 Suitably trained and qualified Research & Development staff in LT	🟥 Population that does not speak the official language(s)	
🟥 Regional or national LT/NLP/ LSP market	Students in language/LT/NLP curricula					Websites with content available in the language	🟥 Policies to provide, maintain and update BLARKs		🟥 Capacity for talent retention in LT	Official, minority and regional languages	
Average socio-economic status	Equity in education					🟥 Web pages			🟥 State of play of NLP/ AI at large	Community languages	
	Inclusion in education					Ranking of websites delivering content in the language			Scientists and researchers working in LT	🟥 Available time resources of the members of the language community	
		🟥 = factors excluded due to missing data				Lables and lemmas in knowledge bases			🟥 Researchers and scholars whose work benefits from the availability of LRs/ LTs	🟥 Civil society stakeholders working on the respective language	
						Language support gaps			Overall research support staff	Speakers attitude	
						E-commerce websites			🟥 Scientific associations or general scientific and technology ecosystems	🟥 Involvement of indigenous people	
									Papers about LT	🟥 Sensivity to barriers that impede the availability of new technology, content and services	
										Usage of Social Media	

Fig. 1: Overview of the contextual factors

Economy	Education	Funding	Industry	Law	Media	Online	Policy	Public Administration	R&D&I	Society	Technology
🟩 **Size of the economy**	🟥 **Academic positions in LT**	🟥 **Public funding available for LT/NLP/AI research projects**	🟩 **Companies developing LTs**	🟩 **Legal status and legal protection**	🟥 **Publicly available subtitled or dubbed media outcomes**	🟥 Digital libraries	🟥 Presecne of local, regional or national strategic plans, agendas, etc.	🟨 **Languages of public institutions**	🟩 **Innovation capacity**	🟨 **Fully proficient speakers**	🟩 **Access to computer, smartphones, etc.**
🟨 Size of the LT/NLP market	🟨 **Literacy level**	🟨 Venture capital available	🟩 Start-ups per year		🟥 **Publicly available transcribed podcasts**	🟩 Impact of language barriers on e-commerce	🟨 Political activity	🟩 Available public services in the language	🟨 Research groups in relevant areas	🟩 **Digital skills**	🟩 **Digital connectivity and Internet access**
🟩 **Size of the language service and translation or interpreting market**	🟩 **Students in language/LT/NLP curricula**	🟩 Public funding for interoperable platforms	🟩 **Start-ups in LT/AI**			🟩 **Wikipedia pages**			🟨 **Scientists and researchers working in relevant aeras**	🟩 **Size of language community**	
🟩 **Size of the IT/ICT sector**	🟩 **Equity in education**					🟩 **Websites with content available in the language**			🟩 **Overall research support staff**	🟨 **Official, minority and regional languages**	
🟥 **Investment instruments into AI**	🟩 **Inclusion in education**	**bold script** = factors which are automatically updateable				🟨 Ranking of websites delivering content in the language			🟩 **Papers about LT**	🟨 **Community languages**	
🟩 **Average socio-economic status**		🟩 = factors with good quality of the data				🟩 **Lables and lemmas in knowledge bases**				🟨 Speakers attitude	
		🟨 = factors with medium quality of the data				🟥 **Language support gaps**				🟩 **Usage of Social Media**	
		🟥 = factors with bad quality of the data				🟨 **E-commerce websites**					

Fig. 2: Classification of the contextual factors

In Figure 2, we show which of these contextual factors can be automatically updated (e.g. via an API of the source, or a script to gather structured information from websites). All information pertaining to the other contextual factors requires some manual processing.

The data per language were then converted into scores that represent whether a language is embedded within a supportive context, ecosystem and climate giving it the possibility to flourish, or whether it may be without political will, funding, innovation and economic interest in the region. The score will, therefore, additionally indicate the probability of a language achieving DLE, given the assumption that a language in an environment with low support will also not be supported from a technological perspective any time soon.

We contend that the DLE metric can accurately reflect the level of LT support for all European languages as an essential requirement for the achievement of full DLE in Europe by 2030. Our preliminary results appear in Section 4.2.3.

3.2 Europe-wide collection of LRTs

To assess the current support of Europe’s languages through LRTs, we need to examine which tools, services, applications, corpora, data sets and lexicons, etc. are actually available for these languages. With more than 30 partners of the project consortium we attempted to systematically collect all existing LRTs for the languages under investigation in the project. As a baseline we used the catalogue of the European Language Grid cloud platform with more than 5000 resources at the time of writing. Together with the various language informants, we managed to identify more than 6000 additional resources, which will soon also be included in the ELG catalogue as proper LRT metadata records. In addition, the ELG catalogue itself will be further enriched by the ELG activity of attaching and harvesting the resources of a number of bigger third-party repositories.

3.3 Language reports

The detailed final results of the ELE metadata collection activity (Section 3.2), a preliminary summary of which is provided in Section 4.1.1, has been used to inform a comprehensive and large-scale review study of the level of support Europe’s languages receive through LT. Conceptualized as updates of the META-NET White Papers (Rehm/Uszkoreit 2012), we have prepared a total of 35 reports on individual European languages (all 24 official EU languages, as well as 11 additional national or regional languages). With the exception of English, German, French and Spanish, 31 of these 35 languages are often considered under-resourced. Each report includes an introduction to the LT field, its main application/research areas and methodologies, general facts about the language, e.g. its status and typology, number of speakers, use on the internet, etc. It also reports the availability of resources based on the combined collection of ELG and ELE resources, the support it receives through dedicated funding programmes and projects, its participation in research infrastructures, and the size of the LT industry in the country/-ies the language is spoken in, etc.

3.4 Online surveys

In order to ensure that our strategic agenda and roadmap has a solid empirical grounding, we collected the views of European users and consumers of LT and also of researchers and developers in the area of LT and AI to consolidate their assessments of the strengths and weaknesses of the field and of the measures that need to be employed so that all European languages benefit from an adequate level of digital provision by 2030. The targeted group of LT researchers and developers comprises: (i) academic and industrial researchers in the field of LT/NLP – beyond pure research, they develop algorithms, pre-commercial LT prototypes, applications and systems; and (ii) innovators and entrepreneurs who commercialize LT to address the needs for digital content analysis and generation, pertinent content transformation and dissemination, as well as for enhanced human-machine interaction. To reach out to this diverse and numerous group of stakeholders, we designed and distributed an online survey addressed to relevant European networks, associations, initiatives and projects. Each respondent was presented with 32 (minimum) to 45 (maximum) questions, depending on their previous answers. The survey was structured in four parts:

- **Part A**: Respondent’s profile, e.g. country, type of organization, LT areas they are mainly active in, participation in networks/associations, etc.
- **Part B**: Language coverage, e.g. languages supported in research, products or services, factors that influence the respondent’s decision with regard to language coverage or support, etc.
- **Part C**: Evaluation of the current situation, i.e. the strengths, gaps and challenges that the European LT community is currently facing.
- **Part D**: Visions for the future, i.e. ideas, predictions and expectations of the LT community about how the LT field as a whole will achieve equal support for all European languages by 2030.

A similar survey was distributed to European LT users and consumers. In addition, we prepared a significantly shortened survey to target European citizens themselves. These stakeholders are often overlooked, but this is ultimately the largest group of users of LT and AI, so it was important to ensure that their views were included. At the time of writing, it looks like we will receive more than 25,000 responses from all countries in Europe, which is very encouraging.

4. Preliminary results

With regard to our goal of achieving DLE in Europe by 2030, our preliminary results first refer to a characterization of the current state in 2022 (Section 4.1) and, second, to the future state in 2030 (Section 4.2).

4.1 The situation in 2022

4.1.1 Europe-wide collection of LRTs

Our systematic collection of language resources, i.e. data (corpora, lexical resources, models) and LT tools/services for Europe's languages (Section 3.2), resulted in more than 6,000 metadata records. This collection has been imported into the ELG catalogue to complement the existing, constantly growing inventory of ELG resources, thus providing information on the availability of more than 11,000 language resources and tools. All languages investigated by ELE are covered, including the official EU languages, non-official, regional and minority languages as well as other European and non-European languages (Fig. 3 and 4).[6] We contend that this collection provides a solid representative basis to investigate the level of technology support for Europe's languages.

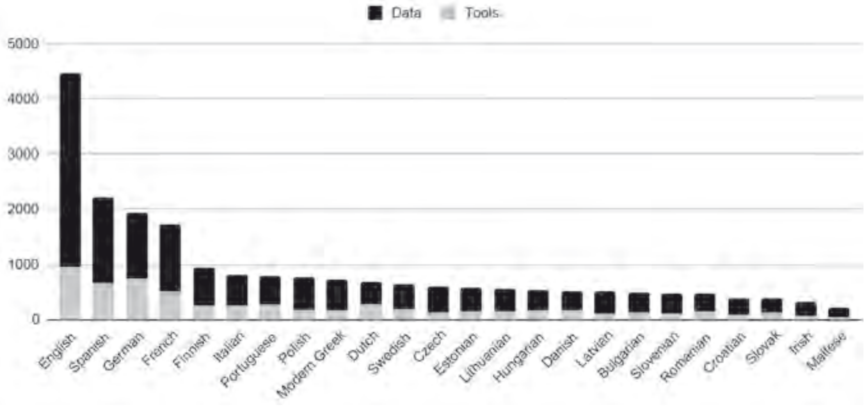


Fig. 3: Number of resources (data and tools) for the official EU languages

[6] Among the languages under investigation by ELE (see https://european-language-equality.eu/languages/), so far no data or tools have been identified for Arberesh, Carpathian-German, Carpato-Rusyn, Cimbrian, Franco Provencal, Griko, indigenous languages in French-Guiana, Jerriais, Meskhetian, Mocheno, Plattdeutsch, Réunion Creole, Romagnol, Southern Italian or Walser.

Figure 3 demonstrates the unsurprising dominance of English, which is represented in 40% of the resources in our collection, followed by Spanish, German and French (each represented in 20%, 17% and 16% of the resources, respectively). A large group of official EU languages occupy the medium ranks, while Irish and Maltese follow in the last positions as the European languages with the most limited technological support. Among the non-EU official languages, two official languages, Norwegian and Icelandic, and four co-official ones, Catalan, Basque, Galician and Welsh, exhibit a noteworthy availability of data and tools. The long tail in Figure 4 provides evidence towards the scarcity of resources for Europe's lesser spoken regional languages, which are practically non-existent in the LT field.

To further investigate whether Europe's languages can be classified in groups in terms of their technological readiness, we considered a set of contextual factors (Section 3.1). One of them is the presence and use of the language in the digital sphere. To measure this factor, we used the number of Wikipedia articles in the language[7] as an indicator, among others. The scatter graph in Figure 5 demonstrates the relation between the amount of data and number of tools in our collection and the number of Wikipedia articles.[8] Four clearly distinct groups of languages emerge from this analysis. English forms a group of its own, as a dominant language, surpassing all other languages by far, both in terms of the number of resources and its digital presence. The second group includes German, French and Spanish. These three languages enjoy a balanced representation in the LT field and on the internet, forming a group of well-supported languages. The third group includes Swedish, Italian, Polish, Dutch and Portuguese, i.e. languages that, despite having an average number of resources, have a sufficiently dynamic digital presence to ensure the availability of raw data that could potentially be transformed into training data for the development of language models and LT applications. The last group includes the remaining languages in Europe, which seem to be poorly supported by LRTs and have a scarce digital presence, which limits their potential for future development. This last group in particular warrants further investigation to reveal possible underlying trends and clusters.

[7] List of Wikipedias: https://meta.wikimedia.org/wiki/List_of_Wikipedias (last accessed 06-11-2021).

[8] The numbers of speakers were mostly derived from online sources, such as Wikipedia and from the language experts in the ELE consortium.

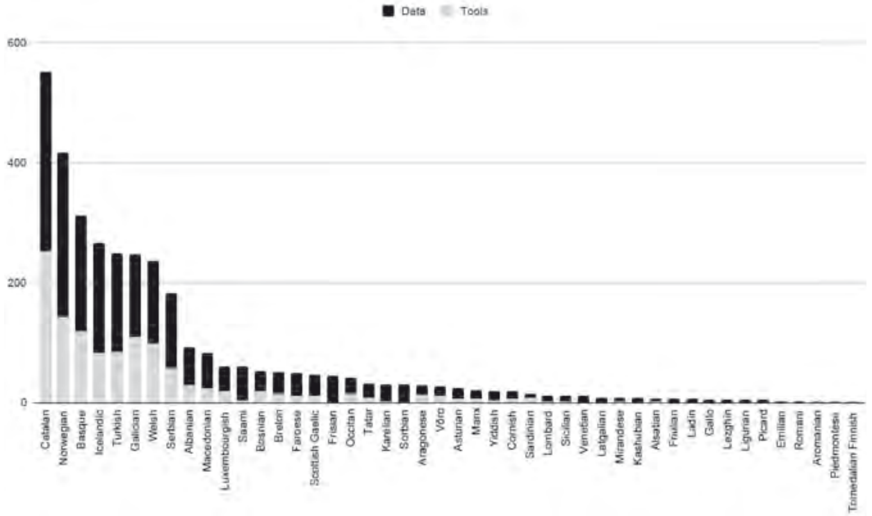


Fig. 4: Number of resources (data and tools) for various non-official EU languages

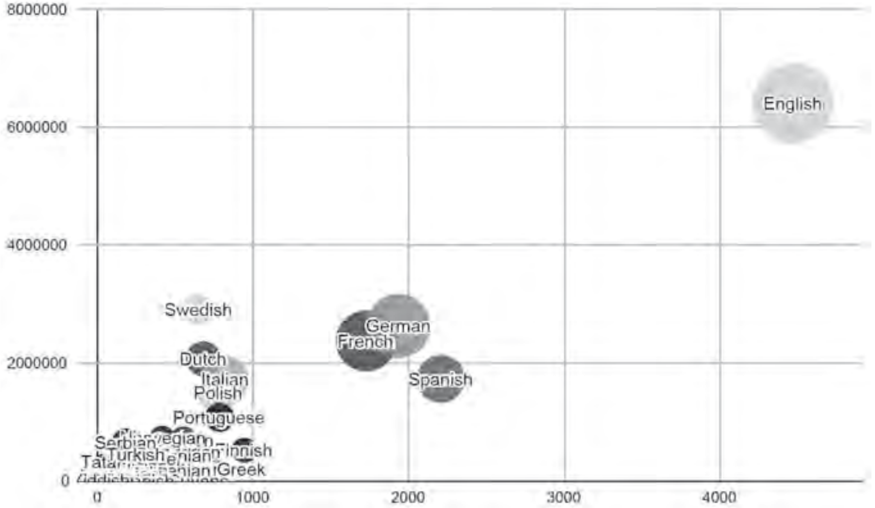


Fig. 5: Number of total resources in our collection vs. number of Wikipedia articles (the size of the circles represents the number of L1 and L2 speakers of the language in Europe)

These findings are largely consistent with those of Joshi et al. (2021), who proposed a taxonomy of languages – “the left-behinds, the scraping-bys, the hopefuls, the rising stars, the underdogs and the winners” – based on resource disparities in the LDC[9] and ELRA[10] catalogues. Like in our study, Joshi et al. (2021) group English, German, French and Spanish in the so-called “winners” group. The main difference compared to our results in Section 4.2.3 is that English is a clear outlier in all statistics based on our collection, thus making it necessary to underline its dominance in the LT world. Nevertheless, this grouping of languages will be further investigated and informed by more contextual factors in future work.

4.1.2 Online surveys

The LT researchers and developers survey (Section 3.4) was online from 17th June 2021 to 18th October 2021. In total, 333 responses were collected. The respondents represent 247 different organizations, of which 74% are research or academic institutions, with the rest being industry practitioners. Geographically the organizations represented are distributed across all EU member states (85% of the respondents) as well as in some other European and non-European countries.

When evaluating the current situation, 88% of the respondents agreed that despite some practitioners declaring a number of applications fuelled by AI as a ‘solved problem’ (e.g. Goodfellow et al. 2016, 473), basic research is still needed. In their open-ended answers, this was specified further, referring to the need to support basic research in linguistics and language modelling, cross-lingual transfer learning and multimodal communication, including speech and sign languages, etc. This was linked to the fact that there are no incentives for research on smaller languages, not only because of the reduced market interest but also because scientific publications reporting on LT-related results for smaller languages are often not considered impactful enough, resulting in a body of scientific literature which is monopolized by results on English. This divide between just a few well supported languages and many smaller ones which are significantly undersupported is further evidenced by the availability of LRs. Low-resource languages will not find their way into industrial processing pipelines or be the topic of large numbers of research publications unless large, high-quality open datasets for these languages become available. In this respect, the role of public funding and procurement was highlighted by the survey respondents, 77% of whom agreed that public procurement is insufficient. Several pieces of feedback noted that smaller languages should rely on public funding to balance the lack of market interest and keep pace in the evolving LT landscape. Among the rest of the most

[9] https://catalog.ldc.upenn.edu/.

[10] http://catalogue.elra.info/en-us/.

frequently mentioned challenges the LT community faces are inadequate recognition of the importance of multilingualism (which 82% of respondents agreed with), the fact that the threat of digital language extinction has not yet made it onto the radars of policy makers or the wider public and competition with and market disruption by non-European big tech companies (82% of respondents agreed with this statement). Finally, it is worth mentioning that the only challenge most respondents do not consider an obstacle is the lack of European talent (54%). The LT community seems to have confidence in the expertise of European human capital as a driving force for the development of LT, although whether this talent pool can be retained in Europe is questionable, especially when one considers the makeup of many of the leading groups worldwide which have a significant European footprint.

4.2 Towards Digital Language Equality in Europe by 2030

The online surveys included a substantial number of responses from the respondents with regard to looking into the future.

Measure/instrument	Avg. Score
• Initiate large-scale, long-term funding programme for European LT development	4.24
• Continuous investment in the Research Infrastructures that support LT	4.23
• Invest in the development of new methodologies for the transfer of resources to other domains and languages	4.05
• Increase availability of qualified personnel on LT and incentives for talent retention	4.03
• Reinforce training & education initiatives, incl. undergraduate & masters programs and vocational training in LT	4.02
• Initiate investment instruments and accelerator programs targeting LT start-ups	3.84
• Public procurement of innovative technology and pre-commercial public procurement	3.79
• Raise awareness of the benefits of the availability of on-line services, contents and products in multiple languages	3.74
• Content accessibility regulations, e. g., multimedia subtitling, readability, dubbing, multilingual content etc.	3.70

Table 1: Average scores (5: very effective to 1: not effective) of the measures and instruments that LT researchers and developers consider effective with regard to LT development towards digital language equality by 2030

4.2.1 Online survey: LT developers

The LT researchers and developers' views and perspectives for future developments towards digital language equality were investigated through a series of closed and open questions.

A critical aspect of the respondents' visions for digital language equality, as brought up in multiple answers, is the availability of resources. By 2030 all European languages should have developed the critical mass of resources that are needed for developing LTs. These include not only raw data but also massive multilingual language models. The issue of data availability was often mentioned in relation to the legal framework for sharing them. Large amounts of data for all languages are expected not only to be available by 2030 but also available for free or at a reasonable cost for both research and commercial purposes. Standardized training and evaluation data for all languages are deemed critical as there is little doubt that shared tasks where such data are made available have significantly helped improve the state of the art in a number of application areas (e.g. *WMT* in MT and Quality Estimation,[11] *SemEval*[12] in Semantics, etc).

In parallel, LT developers are considering working in the coming years towards automated procedures for the construction, annotation and curation of language data, as well as addressing the issue of data bias. Such achievements, combined with continuous work on improving transfer learning methods, are expected to contribute to a situation in which all languages, including small, minority and endangered ones, enjoy technology support and a level of presence in the digital sphere that will ensure their preservation and prosperity.

A shared scientific goal of the LT community is the achievement of Deep NLU by 2030, brought up in numerous responses with various phrasings such as “hybrid intelligence”, “cognitive AI” and “symbolic AI”, etc. All these contributions converge on the description of a future status of LT where the leap from language processing to language understanding has been achieved and seamless human-like interactivity, viable discourse interpretation and ubiquitous natural language interfaces are a reality for all Europeans in their own language. Without wanting to labour the point, however, despite claims to the contrary, we are a long way from achieving these goals.

With respect to the measures and instruments that can be employed to help achieve these goals and realize these visions, the respondents evaluated the effectiveness of a set of proposed measures, as presented in Table 1.

A number of elaborate open answers focused on funding instruments as leverage to help Europe achieve global excellence and leadership in LT. Funding and investments should concentrate not only on the applied (computational) aspects of LT but also on basic research in linguistics and computational linguistics. Support of LR creation and sharing was a constantly recurring issue among the answers we received. With respect to the beneficiaries of funding, a number of survey respondents expressed the opinion that incentives should be provided to language communities that are striving to preserve their cultural and linguistic identities, especially with regard to enhancing a language’s presence on the inter-

[11] E.g. WMT 2021: https://www.statmt.org/wmt21/.

[12] E.g. SemEval 2022: https://semeval.github.io/SemEval2022/.

net. Businesses and industry-research collaborations were noted as an additional target group, and special emphasis was put on limiting bureaucracy in application procedures, which introduces considerable overheads for small companies.

In this context, some respondents perceived the role of national centres of excellence in LT as critically important. Such centres could collect and boost the voices of local players at a national level and increase industry visibility, both nationally as well as at regional and European levels. Apart from designing national research agendas in LT, they should be responsible for the collection, curation, sharing and standardization of language data as well as for employing a European Data Strategy.

Regulatory aspects pertinent to the LT field, in the form of regulations, recommendations or guidelines, were also highlighted. These include, for instance, the adoption of the FAIR principles (Findability, Accessibility, Interoperability and Reuse) in Europe, a revised legislative framework for facilitating the use of language data and the application of data mining techniques for both research and commercial purposes, including guidelines for procurement beneficiaries and public bodies to release their funded/public data, recommendations for both the public and private sectors to provide multilingual websites and for big technology companies to open up their platforms for the lesser spoken languages. The role of the research community is often criticized for its bias towards publications on a small number of the world’s languages. Raising awareness of digital equality issues in the international LT fora and incentivizing Open Access journals and conferences dedicated to less supported languages are among the measures suggested by our respondents to rectify this imbalance.

Raising awareness of the importance of LT for digital interactions and the role of training young LT professionals were mentioned in numerous responses, as were the social dimensions of DLE, which were emphasized by respondents who argued that linguistic and social diversity go hand in hand: the more diverse our society is, the greater the actual need for multi-language resources and technologies. Thus, large-scale policies against racism and discrimination are considered essential. In parallel, engaging minoritized language communities and supporting community building, it is argued, benefit the LT field as it will increase demand for and the impact of LT.

4.2.2 Online survey: LT users

We also collected the views and perspectives of LT users and consumers. The most important finding of this survey is the respondents’ concern regarding the differences in technological support between European languages, specifically the poor technological support of minority, regional and less widely used languages. Various respondents emphasized the need to increase the variety of tools and resources available for these languages. Possibilities include localized social media such as Twitter and personal assistant tools such as Alexa or Siri for languages such as Basque and Catalan. Improved LT support for disabled people is also seen as an important issue. On this topic, survey results reveal the social dimension of LTs that developers should be aware of, and sensitive to, when developing tools and services.

A crucial gap in LTs pointed out by respondents is the limited adaptability of speech technology tools programmed for the most common operating systems such as Android and iOS, which only allow users to use devices developed by Google and Apple, respectively. Thus, software that has been developed by other companies and that supports languages not served by Android or iOS cannot be technically integrated. This observation raises the debate on the need for legal measures to ensure the open and flexible integration of LT services and tools with the most widely used operating systems.

Regarding the provision of resources that would increase the use of language tools for specific languages, the results showed that improved quality coupled with a wider range of tools would increase the use of LTs. When asked about their views on the benefits of improving technologies for the languages they use (including minority, regional and lesser spoken languages), most respondents agreed that LTs can help prevent the disappearance of such languages and increase their numbers of active users. Furthermore, most respondents also agreed that LT can improve communication, even between native speakers, and increase engagement with regard to social, leisure and work activities in their own languages.

With respect to visions for the future, although respondents agreed that in the next ten years there will be higher-quality language tools and a wider range of tools supporting European languages, including minority languages, the results also revealed that many respondents are unsure as to whether, in the next ten years, LT will help prevent the loss of linguistic diversity. Finally, it is worth mentioning that funding to support ongoing work (including that done by freelancers) focusing on the development of tools for minority languages is the main measure suggested by respondents to achieve digital language equality by 2030.

4.2.3 Contextual factors

Following the examination of the range of contextual factors (see Section 3.1), the processing of the data and the development of a scoring method, we were able to calculate scores (normalized to the 0-1 range) for each language which have a strong empirical basis.

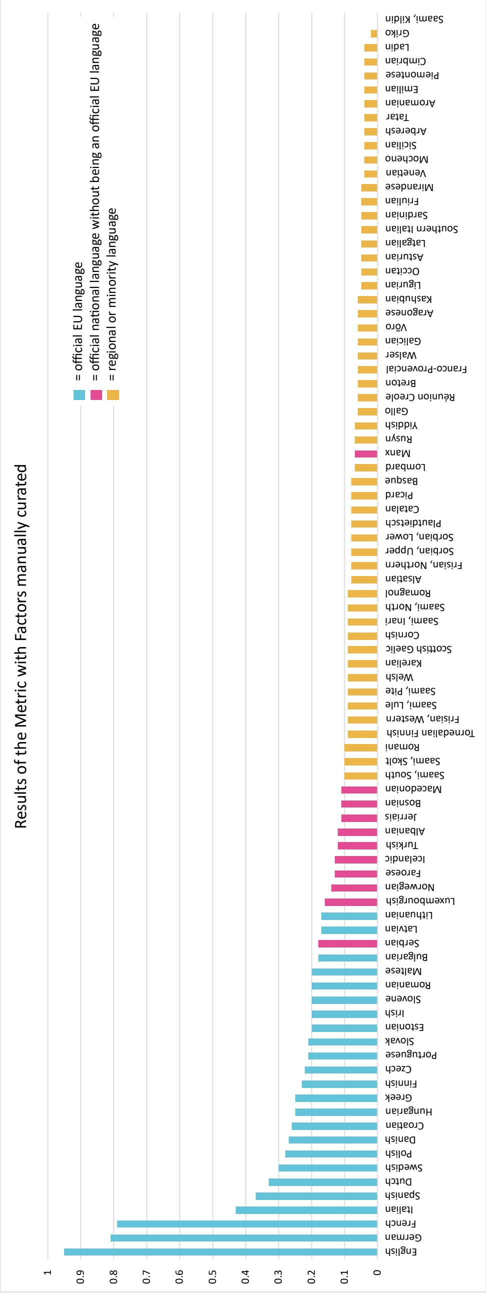


Fig. 6: Results of the 12 manually curated contextual factors

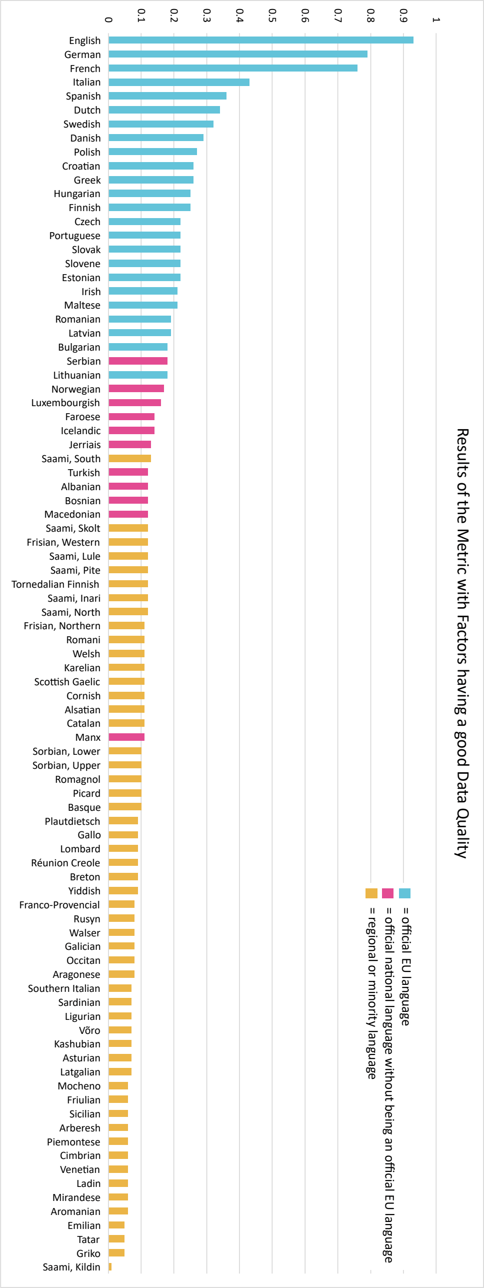


Fig. 7: Results of the 26 contextual factors with good quality of the data

In all configurations that were examined, the top third is dominated by the official EU languages while the regional and minority languages are presented as a long tail to the right. The official national languages which are not recognized as official EU languages appear between the official EU languages and the regional and minority languages. The results of the configuration with 12 selected contextual factors (using four criteria: automatically updatable, having good quality data, not more than 2 factors per class, and a balance between the data types) are shown in Figure 6. Those computed using the 26 factors with good quality data are in Figure 7. Note that each coloured group features instances of single languages from adjoining groups: Serbian in the green group and Manx in the red group.

All configurations clearly demonstrate that English has the best context for the development of LTs and LRs, followed by German and French, with German usually preceding French. Italian and Spanish are in positions 4 and 5. The position of Spanish with a worse score than Italian is caused by only including data from European countries as well as the fact that other languages spoken in Spain are also present in the figures. If data had been included from countries outside Europe, then Spanish, Portuguese, French and English would have had much higher scores given their prevalence in non-EU states. After the five leading languages, variations between the configurations begin to emerge. Mostly, Swedish, Dutch, Danish, Polish, Croatian, Hungarian and Greek are ranked in the upper half of the official EU languages. In some configurations, Finnish also joins this group. The official EU languages with the lowest scores are mostly Latvian, Lithuanian, Bulgarian, Romanian and Maltese.

Among the group of official national languages which are not recognized as official EU languages, Serbian is always the top performer, achieving a score in keeping with the lower-scoring official EU languages, while Manx always appears as a low outlier. Languages such as Norwegian, Luxembourgish, Faroese and Icelandic achieve better scores than Albanian, Turkish, Macedonian and Bosnian. The scores for Jerriais are subject to comparatively large fluctuations, which is why the language is sometimes placed worse and sometimes better.

The regional and minority languages are usually led by Saami, South and Skolt. Depending on the configuration, Tornedalian Finnish, Romani, Northern and Western Frisian and the remaining Saami languages (apart from Saami, Kildin) achieve a score comparable to Saami, South and Skolt. Twenty of the regional and minority languages achieve scores lower than 0.05 in the configuration with 12 selected contextual factors while 31 of the languages obtain scores between 0.06 and 0.1. In the other configurations, the scores of the regional and minority languages are usually higher but with similar differences between the scores of individual languages. Saami, Kildin and Griko are the languages with the lowest scores.

After consultation with our consortium language experts, a number of languages were identified as not being positioned where it was thought they should be in Figures 6 and 7, including Irish, Maltese, Croatian, Latvian, Norwegian, Icelandic, Farose, Jerriais and Manx. Moreover, the regional and minority languages Cornish, Scottish Gaelic, Emilian, Sicilian and most of the Saami languages were rated as not being placed in the correct relative position by at least one of the partners. Overall, this feedback related to 56 out of the 89 languages studied.

We have a number of ways in mind to improve on these results, including adding the vitality status of the language, which is particularly important for regional and minority languages, or adding a factor representing the competition of national languages where more than one official national language exists, and adding statistics on LTs and LRs for languages which are also spoken in countries outside Europe. Nonetheless, as a first cut, we have shown that the DLE metric is a valuable tool on which to base subsequent efforts to measure and improve the readiness of European languages for the digital age, also in the context of the formulation of the SRIIA and roadmap.

5. Summary and next steps

The ELE project is preparing a strategic research, innovation and deployment agenda and roadmap which will provide recommendations on how to achieve digital language equality in Europe by 2030. In this paper, we presented an overview of the project and included preliminary results. Language experts in the consortium have done an extremely thorough job in listing what tools and data exist for a range of European languages, both for official as well as regional and minority languages. A number of surveys have been conducted to elicit responses from a range of stakeholders across Europe. This is very important feedback which will feature in the project's strategic research agenda and roadmap which will clearly outline how digital language equality can be achieved by 2030 for all European languages. Forthcoming results include especially those from the survey which targeted European citizens, with over 20,000 respondents from all over the continent.

In addition, we explained how a range of technological and contextual factors can be used to prime the DLE metric, an extremely useful tool to demonstrate how prepared European languages are for the digital age and what needs to be done to get them to the point where all such languages are digitally equal by 2030. As an extension of this work, we have published our interactive DLE dashboard that makes use of the metadata records available on the ELG platform and provides dynamic visualizations of the DLE metric.

Finally, the strategic agenda and summaries of the main results of the project will be published as a book in the autumn of 2022 (Rehm/Way 2022) and the complete project documentation, including our recommendations, strategic agenda and roadmap, will be handed over to the European Union on schedule in mid-2022. We firmly believe this has the capability of being a game-changer for many European languages which are currently digitally disenfranchised as future funding calls will be geared specifically towards levelling the playing field in this regard.

6. Acknowledgements

Co-funded by
the European Union

This project has received funding from the European Union under Grant Agreement LC-01641480 - 101018166 ELE. Part of the work has also been supported by the ADAPT Centre for Digital Content Technology which is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

References

Ahmed, N./Wahed, M. (2020): *The dedemocratization of AI: Deep learning and the compute divide in artificial intelligence research*. arXiv preprint:2010.15581.

Artetxe, M./Labaka, G./Agirre, E. (2019): An effective approach to unsupervised machine translation. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 194-203.

Bender, E.M./Gebru, T./McMillan-Major, A./Mitchell, M. (2021): On the dangers of stochastic parrots: Can language models be too big? In: *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. New York: Association for Computing Machiney (ACM), 610–623.

Blasi, D./Anastasopoulos, A./Neubig, G. (2021): *Systematic inequalities in language technology performance across the world's languages*. arXiv preprint arXiv:2110.06733.

Bromham, L./Dinnage, R./Skirgård, H./Ritchie, A./Cardillo, M./Meakins, F./Greenhill, S.J./Hua, X. (2021): Global predictors of language endangerment and the future of linguistic diversity. In: *Nature Ecology & Evolution* 6, 2, 163-173. DOI: 10.1038/s41559-021-01604-y.

Brown, T.B./Mann, B./Ryder, N./Subbiah, M./Kaplan, J./Dhariwal, P./Neelakantan, A./Shyam, P./Sastry, G./Askell, A./Agarwal, S./Herbert-Voss, A./Krueger, G./Henighan, T./Child, R./Ramesh, A./Ziegler, D.M./Wu, J./Winter, C./Hesse, C./Chen, M./Sigler, E./Litwin, M./Gray, S./Chess, B./Clark, J./Berner, C./McCandlish, S./Radford, A./Sutskever, I./Amodei, D. (2020): Language models are few-shot learners. In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020*.

Calzolari, N./Bel, N./Choukri, K./Mariani, J./Monachini, M./Odijk, J./Piperidis, S./Quochi, V./Soria, C. (2011): *Final FLaReNet deliverable language resources for the future – the future of language resources. The Strategic Language Resource Agenda*. FLaReNet.

Edunov, S./Guzman, P./Pino, J./Fan, A. (2022): *Teaching AI to translate 100s of spoken and written languages in real time*. https://ai.facebook.com/blog/teaching-ai-to-translate-100s-of-spoken-and-written-languages-in-real-time.

European Parliament (2018): *Language equality in the digital age. European Parliament resolution of 11 September 2018 on language equality in the digital age* (2018/2028(INI). http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

Faisal, F./Wang, Y./Anastasopoulos, A. (2021): *Dataset geography: Mapping language data to language users*. arXiv preprint:2112.03497.

Gaspari, F./Gallagher, O./Rehm, G./Giagkou, M./Piperidis, S./Dunne, J./Way, A. (2022): Introducing the Digital Language Equality Metric: technological factors. In: Aldabe, I./Altuna, B./Farwell, A./Rigau, G. (eds.): *Proceedings of the Workshop Towards Digital Language Equality* (TDLE 2022; co-located with LREC 2022), Marseille, France, 20 June 2022. Marseille, 1–12.

Goodfellow, I./Bengio, Y./Courville, A. (2016): *Deep Learning*. Cambridge, MA: MIT Press.

Grützner-Zahn, A./Rehm, G. (2022): Introducing the Digital Language Equality Metric: contextual factors. In: Aldabe, I./Altuna, B./Farwell, A./Rigau, G. (eds.): *Proceedings of the Workshop Towards Digital Language Equality* (TDLE 2022; co-located with LREC 2022), Marseille, France, 20 June 2022. Marseille, 13–26.

Hassan, H./Aue, A./Chen, C./Chowdhary, V./Clark, J./Federmann, C./Huang, X./Junczys-Dowmunt, M./Lewis, W./Li, M./Liu, S./Liu, T.-Y./Luo, R./Menezes, A./Qin, T./Seide,F./Tan, X./Tian, F./Wu, L./Wu, S./Xia, Y./Zhang, D./Zhang, Z./Zhou, M. (2018): *Achieving human parity on automatic Chinese to English news translation*. arXiv preprint:1803.05567.

Hossain, M. Z./Sohel, F./Shiratuddin, M. F./Laga, H. (2019): A comprehensive survey of deep learning for image captioning. In: *ACM Computing Surveys* (CsUR), 51, 6, 1–36.

Joshi, P./Santy, S./Budhiraja, A./Bali, K./Choudhury, M. (2021): The state and fate of linguistic diversity and inclusion in the NLP world . In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Labropoulou, P./Gkirtzou, K./Gavriilidou, M./Deligiannis, M./Galanis, D./Piperidis, S./Rehm, G./Berger, M./Mapelli, V./Rigault, M./Arranz, V./Choukri, K./Backfried, G./Gómez Pérez, J. M./Garcia-Silva, A. (2020): Making metadata fit for next generation language technology platforms: The Metadata Schema of the European Language Grid. In: Rehm, G./Berger, M./Elsholz, E./Hegele, S./Kintzel, F./Marheinecke, K./Piperidis, S./Deligiannis, M./Galanis, D./Gkirtzou, K./Labropoulou, P./Bontcheva, K./Jones, D./Roberts, I./Hajic, J./Hamrlová, J./Kačena, L./Choukri, K./Arranz, V./

Vasiļjevs, A./Anvari, O./Lagzdiņš, A./Meļņika, J./Backfried, G./Dikici, E./Janosik, M./Prinz, K./Prinz, C./Stampler, S./Thomas-Aniola, D./Gómez Pérez, J. M./Garcia Silva, A./Berrío, C./Germann, U./Renals, S./Klejch, O. (2020): European Language Grid: An overview. In: *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC 2020), Marseille, France. Paris, 3421-3430.

Min, S./Lewis, M./Zettlemoyer, L./Hajishirzi, H (2021): *Metaicl: Learning to learn in context*. arXiv preprint:2110.15943.

Ramesh, A./Pavlov, M./Goh, G./Gray, S./Voss, C./Radford, A./Chen, M./Sutskever, I. (2021): *Zero-shot text-to-image generation*. arXiv preprint:2102.12092.

Rehm, G./Berger, M./Elsholz, E./Hegele, S./Kintzel, F./Marheinecke, K./Piperidis, S./Deligiannis, M./Galanis, D./Gkirtzou, K./Labropoulou, P./Bontcheva, K./Jones, D./Roberts, I./Hajic, J./Hamrlová, J./Kačena, L./Choukri, K./Arranz, V./Vasiļjevs, A./Anvari, O./Lagzdiņš, A./Meļņika, J./Backfried, G./Dikici, E./Janosik, M./Prinz, K./Prinz, C./Stampler, S./Thomas-Aniola, D./Gómez Pérez, J. M./Garcia Silva, A./Berrío, C./Germann, U./Renals, S./Klejch, O. (2020): European Language Grid: An overview. In: *Proceedings of the 12th Language Resources and Evaluation Conference* (LREC 2020), Marseille, France. Paris, 3359-3373.

Rehm, G./Hegele, S. (2018): Language technology for multilingual Europe: An analysis of a large-scale survey regarding challenges, demands, Ggps and needs. In: *Proceedings of the 11th Language Resources and Evaluation Conference* (LREC 2018), Miyazaki, Japan. Paris, 3282-3289

Rehm, G./Uszkoreit, H. (eds.) (2012): *METANET White Paper Series: Europe's Languages in the Digital Age*. 32 volumes on 31 European languages. Heidelberg.

Rehm, G./Uszkoreit, H. (eds.) (2013): *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Heidelberg.

Rehm, G/Way, A (eds.) (2022): *European language equality*. Cham: Springer..

Rosa, R./Dušek, O./Kocmi, T./Mareček, D./Musil, T./Schmidtová, P./Jurko, D./Bojar, O./Hrbek, D./Košťák, D./Kinská, M./Doležal, J./Vosecká, K. (2020): Theaitre: Artificial intelligence to write a theatre play. In: Jorge, A.M./Campos, R./Jatowt, A./Aizawa, A. (eds.): Proceedings of AI4Narratives - Proceedings of AI4Narratives, a Workshop on Artificial Intelligence for Narratives in conjunction with the 29th International Joint Conference on Artificial Intelligence and the 17th Pacific Rim International Conference on Artificial Intelligence (IJCAI 2020), Yokohame, Japan. CEUR Workshop Proceedings 2794, IJCAI, 9-13.

Sanh, V./Webson, A./Raffel, C./Bach, S. H./Sutawika, L./Alyafeai, Z./Chaffin, A./Stiegler, A./Le Scao, T./Raja, A./Dey, M./Bari, M. S./Xu, C./Thakker, U./Sharma Sharma, S./Szczechla, E./Kim, T./Chhablani, G./Nayak, N./Datta, D./Chang, J./Tian-Jian Jiang, M./Wang, H./Manica, M./Shen, S./Yong, Z. X./Pandey, H./Bawden, R./Wang, T./Neeraj, T./Rozen, J./Sharma, A./Santilli, A./Fevry, T./Fries, J. A./Teehan, R./Biderman, S./Gao, L./Bers, T./Wolf, T./Rush, A. M. (2021): *Multitask prompted training enables zero-shot task generalization*. arXiv preprint arXiv:2110.08207.

STOA (2017): *Language equality in the digital age – towards a human language project*. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament. http://www.europarl.europa.eu/stoa/.

Tran, C./Bhosale, S./Cross, J./Koehn, P./Edunov, S./Fan, A. (2021): Facebook AI's WMT21 news translation task submission. In: *Proceedings of the Sixth Conference on Machine Translation*, 205-215.

Wei, J./Bosma, M./Zhao, V. Y./Guu, K./Wei Yu, A./Lester, B./Du, N./Dai, A. M./Le, Q. V. (2021): *Finetuned language models are zero-shot learners*. arXiv preprint arXiv: 2109.01652.

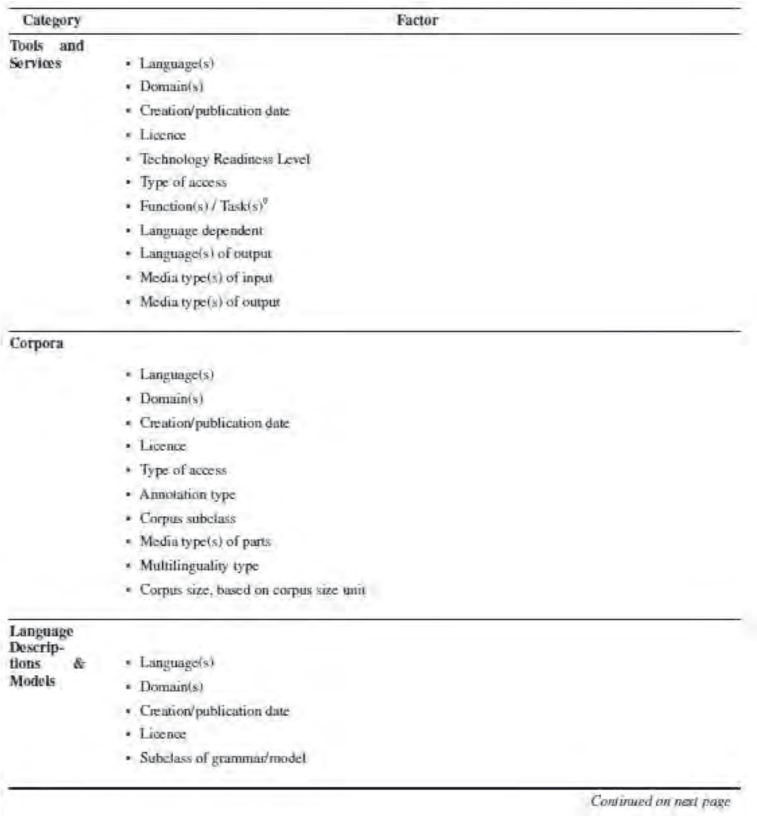
Wu, Y./Schuster, M./Chen, Z./Le, Q. V./Norouzi, M./Macherey, W./Krikun, M./Cao, Y./Gao, Q./Macherey, K./Klingner, J./Shah, A./Johnson, M./Liu, X./Kaiser, L./Gouws, S./Kato, Y./Kudo, T./Kazawa, H./Stevens, K./Kurian, G./Patil, N./Wang, W./Young, C./Smith, J./Riesa, J./Rudnick, A./Vinyals, O./Corrado, G./Hughes, M./Dean, J. (2016): *Google's Neural Machine Translation System: Bridging the gap between Human and Machine Translation*. arXiv preprint:1609.08144.

Ye, Q./Lin, B. Y./Ren, X. (2021): CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, 7163-7189.

Zhang, D./Mishra, S./Brynjolfsson, E./Etchemendy, J./Ganguli, D./Grosz, B./Lyons, T./Manyika, J./Niebles, J. C./Sellitto, M./Shoham, Y./Clark, J./Perrault, R. (2021): *The AI index 2021 Annual report.* arXiv preprint:2103.06312.

Appendix

A.1 Technological Factors



Category	Factor
Tools and Services	• Language(s) • Domain(s) • Creation/publication date • Licence • Technology Readiness Level • Type of access • Function(s) / Task(s)[9] • Language dependent • Language(s) of output • Media type(s) of input • Media type(s) of output
Corpora	• Language(s) • Domain(s) • Creation/publication date • Licence • Type of access • Annotation type • Corpus subclass • Media type(s) of parts • Multilinguality type • Corpus size, based on corpus size unit
Language Descriptions & Models	• Language(s) • Domain(s) • Creation/publication date • Licence • Subclass of grammar/model

Continued on next page

Table 2: Digital language equality – technological factors

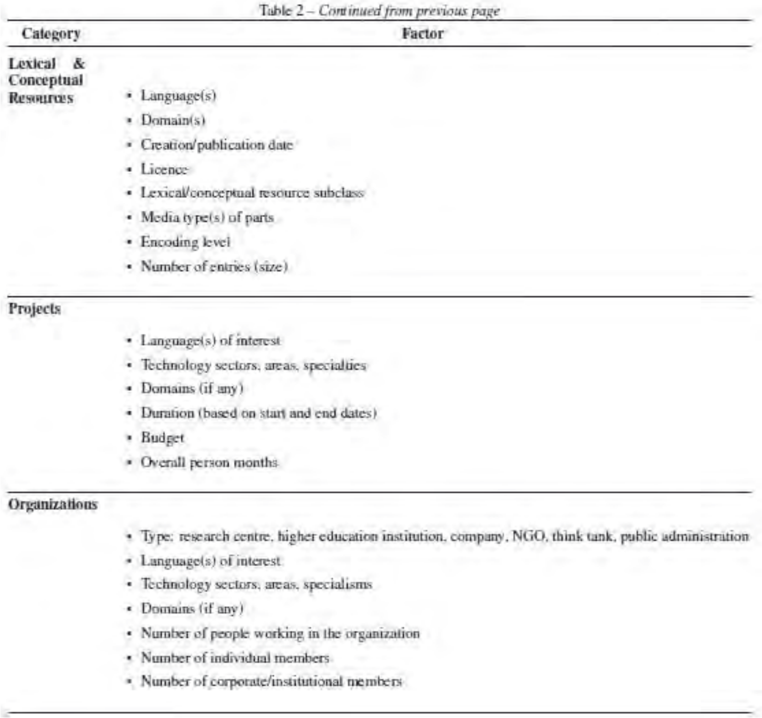
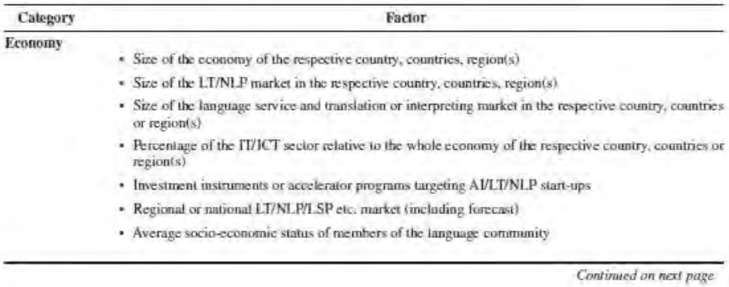


Table 2 – *Continued from previous page*

Category	Factor
Lexical & Conceptual Resources	• Language(s) • Domain(s) • Creation/publication date • Licence • Lexical/conceptual resource subclass • Media type(s) of parts • Encoding level • Number of entries (size)
Projects	• Language(s) of interest • Technology sectors, areas, specialties • Domains (if any) • Duration (based on start and end dates) • Budget • Overall person months
Organizations	• Type: research centre, higher education institution, company, NGO, think tank, public administration • Language(s) of interest • Technology sectors, areas, specialisms • Domains (if any) • Number of people working in the organization • Number of individual members • Number of corporate/institutional members

Table 2: Digital language equality – technological factors (continued)

A.2 Contextual factors



Category	Factor
Economy	• Size of the economy of the respective country, countries, region(s) • Size of the LT/NLP market in the respective country, countries, region(s) • Size of the language service and translation or interpreting market in the respective country, countries or region(s) • Percentage of the IT/ICT sector relative to the whole economy of the respective country, countries or region(s) • Investment instruments or accelerator programs targeting AI/LT/NLP start-ups • Regional or national LT/NLP/LSP etc. market (including forecast) • Average socio-economic status of members of the language community

Continued on next page

Table 3: Digital language equality – contextual factors

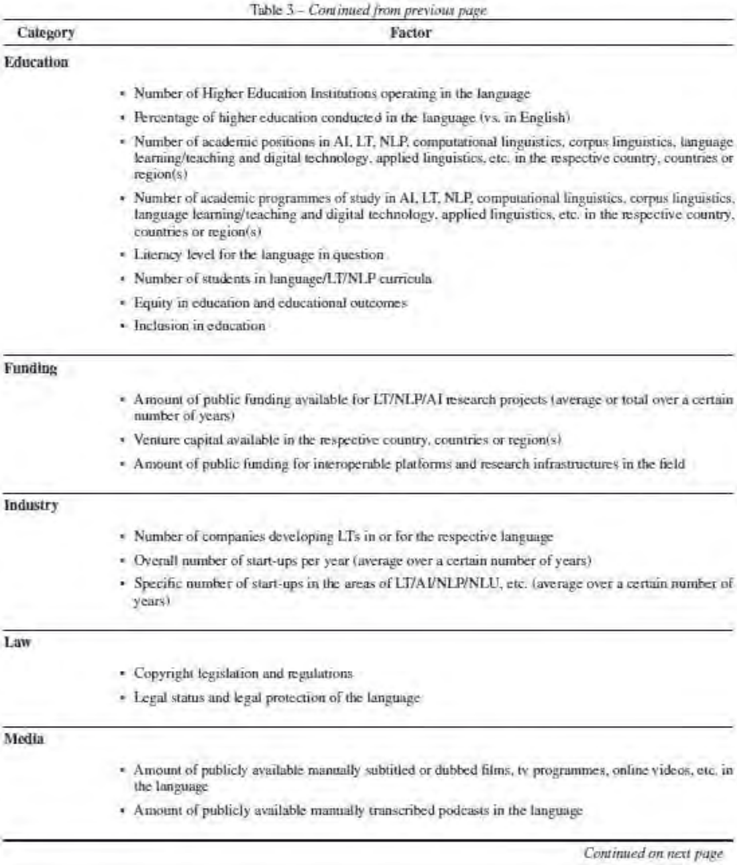


Table 3 – *Continued from previous page*

Category	Factor
Education	• Number of Higher Education Institutions operating in the language • Percentage of higher education conducted in the language (vs. in English) • Number of academic positions in AI, LT, NLP, computational linguistics, corpus linguistics, language learning/teaching and digital technology, applied linguistics, etc. in the respective country, countries or region(s) • Number of academic programmes of study in AI, LT, NLP, computational linguistics, corpus linguistics, language learning/teaching and digital technology, applied linguistics, etc. in the respective country, countries or region(s) • Literacy level for the language in question • Number of students in language/LT/NLP curricula • Equity in education and educational outcomes • Inclusion in education
Funding	• Amount of public funding available for LT/NLP/AI research projects (average or total over a certain number of years) • Venture capital available in the respective country, countries or region(s) • Amount of public funding for interoperable platforms and research infrastructures in the field
Industry	• Number of companies developing LTs in or for the respective language • Overall number of start-ups per year (average over a certain number of years) • Specific number of start-ups in the areas of LT/AI/NLP/NLU, etc. (average over a certain number of years)
Law	• Copyright legislation and regulations • Legal status and legal protection of the language
Media	• Amount of publicly available manually subtitled or dubbed films, tv programmes, online videos, etc. in the language • Amount of publicly available manually transcribed podcasts in the language

Continued on next page

Table 3: Digital language equality – contextual factors (continued)

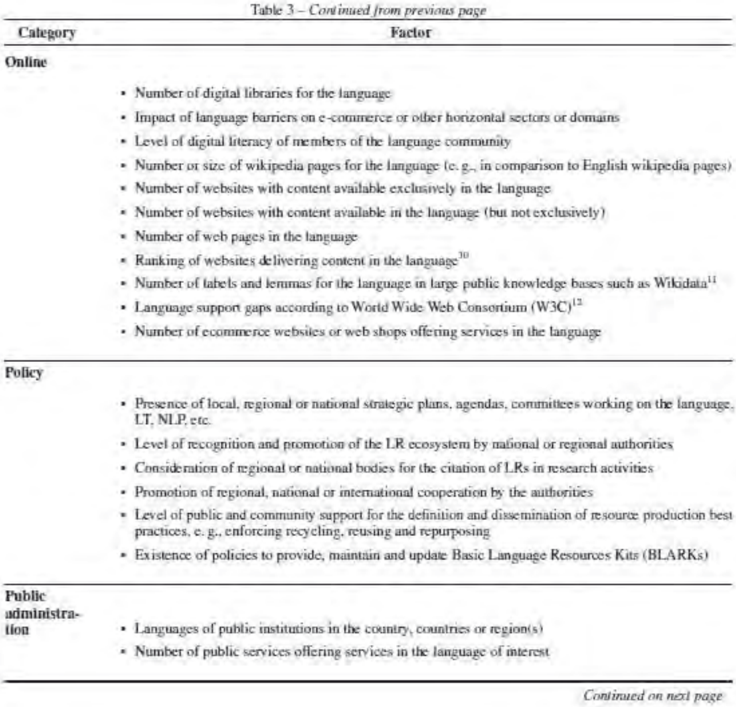


Table 3 – *Continued from previous page*

Category	Factor
Online	• Number of digital libraries for the language • Impact of language barriers on e-commerce or other horizontal sectors or domains • Level of digital literacy of members of the language community • Number or size of wikipedia pages for the language (e. g., in comparison to English wikipedia pages) • Number of websites with content available exclusively in the language • Number of websites with content available in the language (but not exclusively) • Number of web pages in the language • Ranking of websites delivering content in the language[10] • Number of labels and lemmas for the language in large public knowledge bases such as Wikidata[11] • Language support gaps according to World Wide Web Consortium (W3C)[12] • Number of ecommerce websites or web shops offering services in the language
Policy	• Presence of local, regional or national strategic plans, agendas, committees working on the language, LT, NLP, etc. • Level of recognition and promotion of the LR ecosystem by national or regional authorities • Consideration of regional or national bodies for the citation of LRs in research activities • Promotion of regional, national or international cooperation by the authorities • Level of public and community support for the definition and dissemination of resource production best practices, e. g., enforcing recycling, reusing and repurposing • Existence of policies to provide, maintain and update Basic Language Resources Kits (BLARKs)
Public administration	• Languages of public institutions in the country, countries or region(s) • Number of public services offering services in the language of interest

Continued on next page

Table 3: Digital language equality – contextual factors (continued)

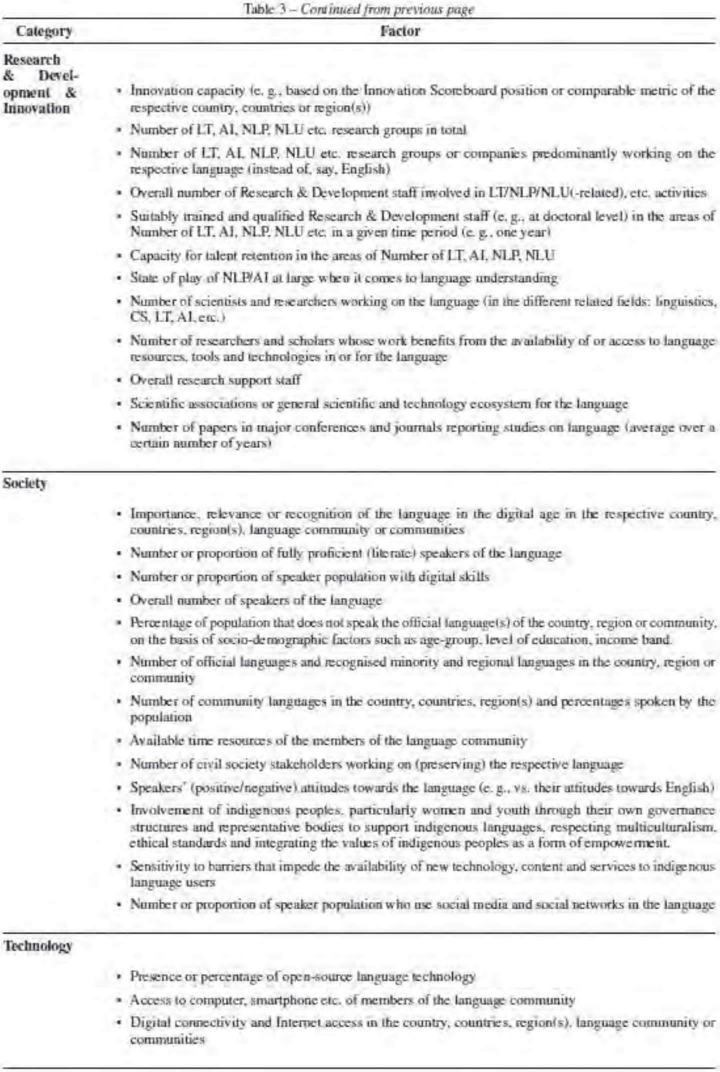


Table 3 – *Continued from previous page*

Category	Factor
Research & Development & Innovation	• Innovation capacity (e. g., based on the Innovation Scoreboard position or comparable metric of the respective country, countries or region(s)) • Number of LT, AI, NLP, NLU etc. research groups in total • Number of LT, AI, NLP, NLU etc. research groups or companies predominantly working on the respective language (instead of, say, English) • Overall number of Research & Development staff involved in LT/NLP/NLU(-related), etc. activities • Suitably trained and qualified Research & Development staff (e. g., at doctoral level) in the areas of Number of LT, AI, NLP, NLU etc. in a given time period (e. g., one year) • Capacity for talent retention in the areas of Number of LT, AI, NLP, NLU • State of play of NLP/AI at large when it comes to language understanding • Number of scientists and researchers working on the language (in the different related fields: linguistics, CS, LT, AI, etc.) • Number of researchers and scholars whose work benefits from the availability of or access to language resources, tools and technologies in or for the language • Overall research support staff • Scientific associations or general scientific and technology ecosystem for the language • Number of papers in major conferences and journals reporting studies on language (average over a certain number of years)
Society	• Importance, relevance or recognition of the language in the digital age in the respective country, countries, region(s), language community or communities • Number or proportion of fully proficient (literate) speakers of the language • Number or proportion of speaker population with digital skills • Overall number of speakers of the language • Percentage of population that does not speak the official language(s) of the country, region or community, on the basis of socio-demographic factors such as age-group, level of education, income band • Number of official languages and recognised minority and regional languages in the country, region or community • Number of community languages in the country, countries, region(s) and percentages spoken by the population • Available time resources of the members of the language community • Number of civil society stakeholders working on (preserving) the respective language • Speakers' (positive/negative) attitudes towards the language (e. g., vs. their attitudes towards English) • Involvement of indigenous peoples, particularly women and youth through their own governance structures and representative bodies to support indigenous languages, respecting multiculturalism, ethical standards and integrating the values of indigenous peoples as a form of empowerment. • Sensitivity to barriers that impede the availability of new technology, content and services to indigenous language users • Number or proportion of speaker population who use social media and social networks in the language
Technology	• Presence or percentage of open-source language technology • Access to computer, smartphone etc. of members of the language community • Digital connectivity and Internet access in the country, countries, region(s), language community or communities

Table 3: Digital language equality – contextual factors (continued)

Per Langgård

How to fight for a digital future – the case of Greenlandic

Abstract

The language technology project that was launched in Greenland in 2005 has attracted quite a lot of attention internationally as one of the few examples of a successful technology project for a lesser resourced language and disproving a hitherto widespread belief that language technology was unrealizable for a language with extreme morphological richness and only a few resources. In this presentation the historical and political background for the project will be outlined and the project's actual progress set out as seen from the viewpoint of the actual developers. A few of the more controversial decisions in the process will be discussed sketchily but the focus will, as far as possible, be kept on observed problems and actual answers to them.

1. Preamble

The presentation in Dubrovnik underlying the present paper was never intended to be very academic and/or theoretical. On the contrary the focus was deliberately kept on empiricism from the viewpoint of a practician developing language technology from within an administrative system not affiliated with a university or any other academic institution.

The present paper will adhere to the same principles. Accordingly, very limited space will be dedicated to methodological considerations and theoretical discussions while the focus, as far as possible, will be kept on observed problems and concrete answers to them.

2. A short introduction to Greenland and Greenlandic

From 1721 to 1953, when it became an integrated part of the Danish kingdom, Greenland was a Danish colony. In 1979 Greenland obtained home-rule, followed by self-government in 2009. On October 1st 2021 56,523 persons lived in Greenland out of whom 89.3 % were born there.[1]

[1] Ethnicity is not recorded in Greenlandic statistics while birthplace is. In spite of the minor uncertainty caused by a small number of children being born to Danish parents in Greenland and a small number of children being born to Greenlandic parents in Denmark, it is comparatively safe to equate birthplace with ethnicity statistically.

Compared to most other small[2] languages, Greenlandic has always been strong and vital with

– linguistic rights never really challenged[3] and constitutionally recognized since the Home Rule Act of 1979. Since 2009, Greenland has been monolingual, with Greenlandic the only official language;
– a standard orthography accepted nationwide since 1861 based on the largest dialect but used in education and administration all over the country. It was replaced by the present (phonemic) standard orthography in 1973. The principle of one national orthography irrespective of dialectal varieties is thus well established in Greenland;
– language policy in local control and never tied to religious or political ideology.

Language is not recorded in Greenland's national register; neither has actual language use and competence been investigated scientifically, but for a rough estimate about half of the population are monolinguals in Greenlandic with no or limited command of Danish L2. About 25 % are believed to be more or less balanced Greenlandic-Danish bilinguals and the rest to have Danish L1 with no or limited command of Greenlandic L2. Greenlandic is thus by all standards a very vital language.

3. Polysynthesis in practice

Greenlandic or Kalaallisut (kal) is the largest dialect in the family of Inuit languages formerly called the Esk-Aleut languages.

Typologically Greenlandic is part of the small group of polysynthetic languages, which, among other characteristic features, include a high level of inflection and a very rich morphology with hundreds of derivational morphemes that combine comparatively freely. A few Greenlandic neologisms will illustrate some of the principles of polysynthesis:

oqaaseq means ‘word’ – in the plural (*oqaatsit*) it means ‘language’;
+PAK is a noun-elaborating morpheme that means ‘several N’. *oqaaserpaat* thus means ‘several words’;

[2] The term “small languages” is used here in spite of the fact that it is considered politically incorrect by some. To me the alternatives are worse, such as the widely accepted term “lesser resourced languages”. Greenlandic, no doubt, has a limited number of speakers and it is correct that the linguistic institutions in Greenland are very limited in size but in other respects, Greenlandic is much better resourced than maybe most other languages. As one example, public and political focus on the language should be mentioned as a very strong resource in Greenland.

[3] The so-called Danification period from around 1950 to around 1975 undoubtedly put quite some pressure on the language, however.

+SUAQ is a noun-elaborating morpheme that means ‘big N’. *oqaasersuaq* thus means ‘a big word’ and *oqaaserpassuit* (oqaaseq+PAK+SUAQ) ‘very many words’;

-LIRI is a verbalizing morpheme that means ‘deal with N’. *oqaasileri-* thus means ‘work with language’, *oqaaserpaleri-* (oqaaseq+PAK+LIRI) means ‘deal with a number of words’, *oqaasersualeri-* (oqaaseq+SUAQ+LIRI) means ‘deal with a big word’ and *oqaaserpassualeri-* (oqaaseq+PAK+SUAQ+LIRI) means ‘do language technology’;

+NIQ is a nominalizing morpheme that forms abstract verbal nouns. *oqaasilerineq* thus means ‘linguistics’ and *oqaaserpassualerineq* means ‘language technology’.[4]

As can be seen, one stem combined with four out of several hundred derivational morphemes generates 12 new stems. If we include inflectional morphology these 12 stems alone will produce more than 3,000 wordforms that all combine freely with about 50 enclitic morphemes generating more than 150,000 individual wordforms.

The rich morphology is a challenge for Greenlandic language technology but, as a matter of fact, a minor problem compared to the syntax problems caused by inderivation[5] and the fact that a number of features like gender, definiteness and tense have no immediate morphological manifestations.

Polysynthesis is a challenge for Greenlandic language technology but not an unsurmountable one as a concrete parsing example will demonstrate. In three different sentences, the same wordform *kusanartumik* (beautiful) has three different syntactic functions as (1) an adnominal argument to an inderived object, (2) an adverbal argument to a main verb and (3) an adverbal argument to an inderived verb inside a noun:[6,7]

“<kusanartumik nuliaqarpoq >” *He has a beautiful wife*
“kusanar” TUQ N Ins Sg @i->N #1->2
“nuliaq” QAR V Ind 3Sg @PRED #2->0

[4] Note that also *oqaaserpalerineq* and *oqaasersualerineq* are well-formed words.

[5] The process when a stem after derivation forms part of a new stem of another word class but maintains its original syntactic features. See Langgård (2002) for a thorough introduction to this issue.

[6] A number of secondary tags for use with higher level analyses have been stripped from the examples for clarity.

[7] The tags in the example: TUQ is a nominalizing derivational morpheme ‘one who Vb’; N is a ‘noun’; Ins is the ‘oblique case instrumentalis’; Sg is ‘singular’; @i->N ‘”adjective” to an inderived object’; QAR is a verbalizing derivational morpheme meaning ‘have N’; V is a ‘verb’; Ind is ‘indicative mood’; 3Sg is ‘subject’s person is 3. sing.’; @PRED is ‘main verb’; @ADVL> is ‘adverbial pointing right’; @i-ADVL> is ‘adverbial to inderived verb pointing right’; Abs is the ‘case absolutive’; 1Sg is ‘subject 1. sing.’; 3SgO is a ‘verb inflected for 3. sing. object in the transitive verb’; #n->n are dependencies.

“<kusanartumik oqaluppoq >” *He talks beautifully*
“kusanar” TUQ N Ins Sg @ADVL> #1->2
“oqalup” V Ind 3Sg @PRED #2->0

“<kusanartumik oqaluttoq naapippara >” *I met somebody talking beautifully*
“kusanar” TUQ N Ins Sg @i-ADVL> #1->2
“oqalup” TUQ N Abs Sg @OBJ> #2->3
“naapip” V Ind 1Sg 3SgO @PRED #3->0

As can be seen, the Greenlandic parser has the capacity to automatically distinguish between the different grammatical structures and tag all words adequately.

4. The Greenlandic language technology project – preconditions

Deliberate language planning has always been part of language policy in Greenland. Before 1959, when Landsrådets sprog- og retskrivningsudvalg[8] (the first government institution for language) was introduced, language policy was not explicitly set out in the colonial and early post-colonial administration of Greenland but there can be no doubt that the laissez-faire attitude towards Greenlandic clearly included much respect for the native language of the colony. For instance Greenland’s first nationwide newspaper, Atuagagdliutit[9] founded in 1861, was monolingual in Greenlandic. It was printed in Nuuk and distributed free of charge explicitly in order to strengthen the orthographical standard of 1851 (Oldendow 1957).

From around 1990, when grammar and spell checkers started to be used regularly in Danish and English word processing programs, requests for comparable tools for Greenlandic were occasionally heard and a few attempts were actually made to construct Greenlandic spell checkers based on word lists around the turn of the century. With a detection rate as low as 20-25% they were useless but the wish for language technology to support the vulnerable Greenlandic language slowly started to grow, although it was generally considered an impossible endeavour for a small language. It should be noted that such attitudes were normal among laypeople and language professionals alike.

This discourse began to change in 1999, the beginning of Greenlandic language policy and language planning in its current form, when an academic secretariat

[8] This can be translated ad hoc by “The local parliament’s committee on language and orthography”.

[9] *Atuagagdliutit* literally means ‘reading matter given away [for free]’. As a curious but interesting side note, *Atuagagdliutit* was the world’s first newspaper with colour illustrations.

for the parliament's three[10] standing committees on language was established with a staff of two. The new institution was later given its present name, Oqaasileriffik/ The Language Secretariat, and has since grown to its present staff of eight.

Already before 1999 there was public and political awareness of language technology as a support for the vulnerable Greenlandic language and there were a few attempts to produce concrete technology. Especially Henrik Aagesen's morphological parser, Qimawin (Aagesen 2004), should be mentioned as a fine example of mature language technology provided by an independent researcher at an early stage. Unfortunately, Qimawin never got the attention it deserved academically and never came into widespread use.

As soon as Oqaasileriffik had been set up, it focused on compiling basic resources and adapting an existing grammar of Greenlandic to prepare it for machine readability. By 2005 the lexical resources and grammatical description had reached a level that made it possible to start up the language technology project on a more ambitious scale.

5. The Greenlandic language technology project – expected and observed obstacles in the run up to the project's launch in 2005

Oqaasileriffik almost immediately realized that the real problems facing the establishment of a language technology project in a minority society with rather traditional and conservative values were very different from the ones one could expect to have to face. While Oqaasileriffik expected typological questions and technological problems to be the main challenges, it soon became clear that a number of attitudinal problems were much more severe and had to be faced and addressed before the project could be launched:

- In Inuit societies, the primary opinion formers in relation to traditional culture including language are the elders. In their opinion language technology was unnecessary outsiders' technology;
- Polysynthetic Greenlandic deviates far too much from languages traditionally associated with language technology. In addition, the scarcity of training data is expected to render any Greenlandic projects undoable;
- Language technology presupposes a staff of specialized computational engineers and an advanced state of IT infrastructure. Neither exist in Greenland and outsiders' support is not an option since almost no non-Greenlanders speak Greenlandic;
- Language technology is prohibitively expensive;

[10] To be precise only the language board and the place names' committee were actual committees while the decision-making body concerning personal names was a rather independent work group affiliated with the bishop's office.

– Should Greenland succeed – against all odds – it will be of no use anyway since the tech giants will never add local tools to their applications for economic reasons.

So while Oqaasileriffik established the process of compiling basic resources, at the same time it had to invest a considerable amount of energy in dialogues with society and in public debates about modern language planning. Fortunately, access to the media is rather open in the small society as are possibilities to deliver public presentations. Both were extensively exploited at the same time as Oqaasileriffik, with the help of small grants for limited projects, was able to publish small applications paving the way for the funding of future, more ambitious projects while attempting to design them in such a way that they could be economically acceptable for funding by Greenlandic public means.

Apart from struggling with the inveterate belief that language technology for Greenlandic is impossible for typological reasons, one other attitude drew much energy from creative work. Conservatism and the high level of respect for elders well known almost everywhere in first-nation societies proved to be major obstacles for a qualified dialogue with society. For Oqaasileriffik to be taken seriously and to pave the way for future funding, the fact that “real” Greenlandic is much more than the elders’ sociolect as well as their purism had to be addressed directly. For Greenlandic to survive in the modern world, society had to learn to accept the fact that any language must be able to adapt to hitherto unknown registers and domains. A language used exclusively for local affairs in the past will not survive long.

After a few years as outlined above, the compilation of basic resources had reached a certain size and a new public discourse ready for a language technology project seemed to have emerged so a project constructing the first Greenlandic finite-state transducer was launched in 2005, when Oqaasileriffik received a small grant to relieve one staff member of other duties and got a head start because of generous start-up support from several Nordic universities. Especially Giellatekno in Tromsø directly facilitated the project, including extensive, private teaching of Oqaasileriffik’s staff. Without Giellatekno’s support in the project’s early days, Greenlandic language technology would not have been anyway near its present status.

After a year’s work, the first finite-state automaton was mature enough for a spell checker and a few small online tools to be constructed. The spell checker had a modest detection rate of around 80 % and the tools were rather primitive but they were enthusiastically received by society.

They concretely proved that Greenlandic language technology made by local staff is doable, which, looking back, might have been its most important impact.

Over the next few years, the automaton was debugged and expanded with the help of small grants interspersed with periods without funding. This changed dramatically in 2011 when a €400,000 grant from the Danish Velux foundation enabled Oqaasileriffik to expand the tagger project into a parser project and hire two BA students for in-house training.

From the very beginning it was obvious that training was the key to success and had to be an integral part of the project since there was never the option to pick qualified staff "off the shelf". The study of language technology was not offered anywhere in Denmark in those days.[11] Furthermore, it was next to impossible to raise interest in language technology in the younger generation and to attract students. During a nation-building era, cultural studies, history and other academic disciplines that can be immediately related to a reborn identity as a non-European Inuk were in very high esteem while it proved difficult for the newly established university to rouse students' interest in "European" studies like formal linguistics and computer science.

The training aspect is crucial and to a high degree explains why the 2011-grant turned out to be the paradigm shift it actually was. So to secure the project's future we had to accept in-house training although it was very time consuming for senior staff.

6. Summing up the challenges and actions taken to answer them

The Greenlandic language technology project is believed to have achieved much better results than almost all other LT projects for very small languages.[12] In Greenland politicians and language administrators are convinced that this is explained to a large extent by the fact that Greenlandic language policy has been consistent, also in situations where the public has been critical of elements of the policy. For instance Greenland has always had only one robust national orthography in spite of rather deviating dialects. Especially among the 3,000 speakers of East Greenlandic this policy is resented by many but the political demand for only one standard has never been seriously challenged. In many minority societies with a more permissive view on dialects, Greenland's one orthography policy is often questioned but the parliament considers standard orthography to be an important tool in preserving Greenlandic.[13]

[11] There were options outside Denmark and a few Greenlanders were actually involved in such programs but the challenges for a Greenlandic speaking student with limited Danish L2 and less English L2 proved to be too prohibitive for us to exploit that option. Also the economic aspect should be mentioned. It is extremely costly to send Greenlandic students to universities outside Denmark.

[12] The Giellatekno project for Samic languages is one important exception.

[13] Whether this is the case or not shall not be debated here but it should be pointed out that neither East Greenlandic nor Inuktun in the northernmost part of West Greenland is critically endangered after almost 150 years without local orthographies.

Another factor that is believed to be important in keeping Greenlandic strong is the fact that Greenland, unlike Canada and Alaska for instance, did not put cultural control in the hands of Elders' Councils or the like. Instead comparatively young ministers of culture and directors at Oqaasileriffik have counterbalanced the elders' purist agenda and broadened public opinion about "correct" language.

Another question should be addressed in this context, namely the degree of ambition. In most small languages the criterion for success is keeping the local language alive in relation to local matters while leaving all non-traditional matters like technical terminology, higher education, foreign trade, etc. to be handled by the nearest majority language. In Greenland this is not an option. Neither inside nor outside parliament are voices to be heard advocating diglossic approaches to technical terminology, for instance. Even that must be localized.

Language policy is explicitly set out in the Self Government Act of 2009 to be unrestrictedly monolingual in Greenlandic. A language policy as ambitious as outlined here is, of course, strenuous everywhere in language administration and education but still the policy is believed to have contributed considerably to the healthy state of the Greenlandic language over the years.

7. Conclusions

Greenlandic is extremely vital in comparison with other small languages. At Oqaasileriffik, it is our firm belief that the rather restrictive, albeit not puristic, language administration pursued over many years has played an important role in ensuring that Greenlandic remains alive and healthy.

The Greenlandic language technology project is an important part of the overall picture as its success depends to a large extent on the fact that it evolved on the basis of a robust standard variety and that language technology in turn reinforces said standards.

Once this starting point of limited permissiveness in both status and corpus planning in Greenland has been established, a few principles and experiences should be mentioned that are believed to have been important in keeping the project alive and growing for so many years.

The project has to be anchored locally. As mentioned earlier, we received much support from Nordic colleagues in the early stages of the project. Fortunately, this support never came in the shape of ready-to-use programs developed outside Greenland. Instead it had the shape of helping to help oneself. Therefore the overall project design as well as all of the tools has been produced locally in Greenland. It should be observed that this does not imply an unwillingness to reach out for help from outside. On the contrary, the small and fragile milieu of Greenlandic language technologists is almost constantly in need of much help – and is lucky enough to get most of what is asked for. But there are important preconditions to the nature of the help asked for. Only solutions that can be maintained and updated locally by local staff are welcome.[14] This includes the necessity of importing only know-how at a level of abstraction that is viable for the local competence and local education of the local workforce.

The project has to be more ambitious than probably any other language technology project for a language with resources comparable to the Greenlandic ones. The unavoidable fact that a language is a language no matter how few speakers it has is not a question of degree. Accordingly, attempts to develop resources for a variety of any language exclusively for local use in connection with local affairs is not enough. In our globalized world, even small languages need to address unknown topics and unfamiliar domains as much as major languages do. Therefore the Greenlandic language technology project deliberately included "difficult stuff" like technical terminology, neologisms and the like almost from the very beginning.

Only technology that is multifunctional and versatile is viable. Greenland has very limited resources both in terms of manpower and money. One such nonexistent resource is a manned institution for NLP using mainstream techniques like machine learning, AI and the like. Accordingly, Greenland must rely on other technologies. Rule-driven technology is an approach Greenlanders can depend on without relying on a foreign workforce because the technology puts limited demands on computational know-how and because Greenlanders are the ones who know the language intimately. It is also a very versatile technology. Once basic lexical resources have been compiled and a tagger and a parser developed, this one set of resources will suffice to construct a number of applications and tools including spell checkers, grammar checkers, and L2 material. It will also take an MT project far if paired with a glossing device.

Permissiveness is a luxury most minority languages cannot afford. This postulate is extremely controversial but Greenlandic decision makers are convinced that there is no alternative if a vulnerable language like Greenlandic should survive for future generations.

As noted repeatedly above, human and economic resources for the Greenlandic language are extremely limited. After the introduction of Home Rule in 1979, the overall situation for the administration of the Greenlandic language obviously improved a lot. Funding has improved dramatically and after establishing an institute for Greenlandic language when Ilisimatusarfik/Greenland's University was

[14] There is one important exception. A number of years ago Oqaasileriffik bought a larger application from outside that has proven to be too technical and complex for local competences and has tied Oqaasileriffik to some legal restrictions which cannot be controlled locally. That application is still running and will do so for a number of years until an alternative developed and controlled by Greenlandic human and economic resources can be established.

founded,[15] a small group of Greenlandic-speaking language professionals has emerged making it possible for Oqaasileriffik and the university to fill a dozen or so positions for the administration of Greenlandic and teaching Greenlandic at university level.

Still, although this recent development is very positive, the fact remains that the needs are many and extensive, leaving Greenland in the sad position of efforts put into activities outside a narrow core of daily obligations and immediate political demands for new tools and facilities will inevitably drain resources from the core activities.

So out of necessity rather than inclination, Oqaasileriffik has only rudimentarily included dialects, dialectisms, and varieties such as Facebook-Greenlandic in the basic resources and applications which have been developed recently.

In terms of controversiality error correction is in a league of its own. To most fellow language technologists, applications in general should not always expect correct input. Instead, the programs should deal with typos and other inaccuracies, including dialectisms, in a clever way and process input seamlessly as if the input was given in the expected standard. Greenlandic politicians have explicitly asked Oqaasileriffik not to include error correction to any large degree in our language technology project for pedagogical reasons. A high level of L1 language awareness is, namely, considered important for future vitality and error correction is believed to be counterproductive to this political aim.

Accordingly, the Greenlandic language technology project is basically prescriptive apart from neologisms and morphological reductions of a certain frequency as well as grammaticalization at all descriptive levels. Such natural developments are considered in the present work.

The staff at Oqaasileriffik does not, a priori, see such a political demand as an unjust restriction to their work. The bottom line is that the standardization and prescription that have prevailed in Greenland as far back as records go appear not to have been harmful to the vitality of the language. On the contrary, the dramatic decline in fellow Inuit languages in Canada and Alaska that in many respects are comparable with Greenlandic but where dialectal diversity has been a priority in language policy definitely does not go unnoticed.

We know, of course, that no causality can be postulated exclusively on the basis of this observation but the fact remains that Greenlandic is vital and healthy in the realm of present language policy and that this is a fact we feel we have to consider.

[15] Ilisimatusarfik officially became a university in the parliamentary act of May 9th, 1989, but before that a BA in Greenlandic culture including some focus on language had been an option at the university's predecessor, the Inuit Institute in Nuuk, since 1984.

8. The future for the Greenlandic language technology project

Oqaasileriffik expects the present development to continue and expand in the near future. The basic resources generally have standards which are high enough to develop a wide range of tools and applications as well as to improve existing ones. Likewise funding seems to be secure at least at the present level in the immediately foreseeable future and hopefully beyond.

Funding has actually improved in 2022 with a new chair for a terminologist created in this year's Finance Act and a substantial grant received from the Danish parliament for a private entrepreneur to improve and expand a language technology based Greenlandic L2 system. It is expected that the Danish grant will create much synergy with the projects at Oqaasileriffik.

Apart from matured resources and improved applications, the years to come will see new ones especially in the fields of technical terminology and pedagogical materials for Greenlandic L2. Furthermore high priority will be given to English in Greenland. English resources are scarce and the need for adequate teaching materials at school which do not presuppose Danish as a bridge to English is great as is the general need for modern dictionaries between Greenlandic and English.

One aspect, though, of English in Greenland calls for special attention, namely the great impact of English on Greenlandic via the tech giants that is rapidly increasing everywhere in Greenland after the sea cable laid in 2008 made general access to the internet better and cheaper.

No valid information on the phenomenon is available but quite a number of personal observations and calls from worried parents about Greenlandic children communicating with other Greenlandic children in pidgin-style English suggests that the problem is growing. The primary sources for this kind of English are allegedly YouTube and gaming but extensively used non-localized applications like Google, MS Office and the major operating systems by the adult population are expected to add to the picture.

This present development might be the biggest threat to Greenlandic ever experienced but no one knows what can be done about it. Extensive localizing might reduce the dangers but nothing like that seems to be on the tech giants' cards.

Oqaasileriffik has tried hard to get into contact with the tech giants about the problem but nothing approaching a dialogue has come out of that. Most of the correspondence is simply ignored and on other occasions, we get what seems to be robot-generated reactions that do not address the problems written about at all.

At the moment several initiatives for a working group under the auspices of the Nordic Council of Ministers are in the making but it is not yet possible to predict whether they will be more successful than earlier attempts by Oqaasileriffik.

Still, the clock is ticking and reports on English affecting children's Greenlandic mother tongue are growing more numerous all the time so idling is not an option for Greenland. The impact of English on Greenlandic is already a fact and nothing implies that this will diminish in the foreseeable future. To prepare for such unavoidable bilingualism in cyberspace lots of work must be done soon. This includes the production of Greenlandic-English MT to render localization a possibility[16] as well as serious refinements of Greenlandic writing aids and lexical resources to make the mother tongue competitive towards English L2 also in technical domains, just to mention two of the many achievements needed

That is all very far away but standing still is going backwards so something must be done. Added to this is the fact that all achievements on route for that goal will improve the quality of tools and lexical resources that all Greenlanders will need access to for Greenlandic to survive in the shadow of English L2.

References

Aagesen, H. (2004): *Qimawin: en grønlandsk ordkløver og samlet ordbog : program til PC med Windows95 eller nyere*. Nuuk: Atuagkat.

Langgård, K. (2002): Inderivation in Greenlandic. In: Nedergaard Thomsen, O./Herslund, M. (eds.): *Travaux du Cercle Linguistique de Copenhague Vol XXXII. Complex Predicates and Incorporation a Functional Perspective.* Copenhagen: C.A. Reitzel, 67–119.

Oldendow, K. (1957): *Bogtrykkerkunsten i Grønland og mændene bag den: en boghistorisk oversigt*. Copenhagen: F.E.Bording.

Statistics Greenland: stat.gl/dialog/main.asp?lang=da&version=202104&sc=BE&colcode=o

[16] This assumes that the tech giants, in time, will open up their applications for third-party software like Greenlandic MT.

Trond Trosterud

Normative language work in the age of machine learning

Abstract

Neural nets have, during the last few years, given us both an improved Google Translate, better search algorithms, better speech technology and doubtless many other things. The approach dominates current language technology to the extent that no other approach is visible. Being data driven, the hidden assumption behind this approach when used in proofing tools is that the language is used correctly in the text material, in other words, *usage equals the norm*. Although this approach is able to provide useful help for the largest languages, it leads to some serious problems. For indigenous and often also for other minority languages, the assumption does not hold. The written norm is weakly established and cannot be reliably found in usage. For normative bodies responsible for defining the written norm of a given language, usage-based proofing tools will not be able to implement the explicit norm they have defined. The present article discusses the current trend within proofing tools and looks at some alternatives.

1. Introduction

When politicians ask, language technologists answer that all they need is more data, i.e. they need a Language Bank. When constructing language tools, their preferred method is the one that **trains** the computer. The use of AI within the field of planning and implementing written norms thus increasingly equates to adding more text to the tool and hoping for the best.

This works for language societies where there is much text available, the language does not have dynamic compounding and correct forms clearly outnumber incorrect ones. However, for most languages, these assumptions do not hold.

In order to understand the role of text and explicit norms in language planning we must understand the current trends of language technology, which no doubt include the trend of machine learning from Big Data. Language technology applications are, to an increasing degree, constructed with the help of large data collections by large companies whose main focus is outside language technology. These companies will never have national language planning high on their agenda. Their optimal scenario seems to be data-driven language technology with as few philologists as possible, which is easy to roll out for new languages and with minimal additional costs for each new language. The focus is on the customer, who did not buy proofing tools but got them “for free” when buying something else, and not on the language community as such.

2. Proofing and dynamic compounding

Dynamic compounding is found in Europe in the area between English, Slavic, and Romance, i.e., it covers the Germanic, Finnish and Saami language area. In these languages, compounds like *reindeer husbandry agreement negotiations* are written as one word, with non-trivial distribution of internal morphology (the Norwegian suffix -*s*- is historically a genitive suffix), as shown in (1):

(1) *reindriftsavtaleforhandlingar* (*Norwegian*)
rein-drift-s-avtale-forhandling-ar
reindeer-operation-COMPSUFF-agreement-negotiation-PL.INDEF

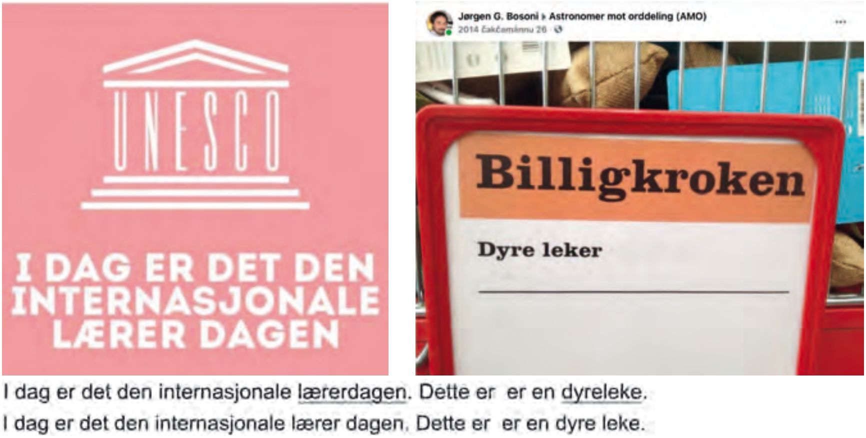
poronhoitosopimusneuvottelut (*Finnish*)
poro-n-hoito-sopimus-neuvottelu-t
reindeer-GEN-operation-agreement-negotiation-PL

reindeer husbandry agreement negotiations

The compounds in (1) are lexicalised, but also ad hoc neologisms like Finnish *yhdyssanakeskustelufoorumi* (“compound word discussion forum”) are perfectly fine.

Now, the question is how this may be handled in a spellchecker. There used to be three ways of making a spellchecker: the wordform list approach, the stem + affixes approach and the grammatical approach. The wordform list approach is good for languages with no or almost no morphology, like most Polynesian languages or even English. The stem + affixes approach is a good fit for languages with regular suffixation, such as Turkish or the Uralic language Komi. In the grammatical approach, stems and affixes are paired with lexeme and grammatical properties and subsequently combined with a model dealing with morphophonological processes. This spellchecker is good for languages with complex morphology, like the Saami languages or Finnish.

The two first methods dominated until the 1990s, and still do in many contexts. What they have in common is that they do not handle dynamic compounding. As a result of this, erroneously split compounds became common with the introduction of computers and spellcheckers during the late 1980s. The two examples in Figure 1 are taken from a Facebook group devoted to making fun of such errors. The first example, celebrating international teachers’ day, shows that (people advertising for) teachers also make these mistakes. The second example shows that the basket containing cheap commodities, *Billigkroken*, does not contain “animal toys" (*dyreleker*), as intended, but instead contains "expensive toys" (dyre leker). This error type may certainly be due to influence from English, but what is relevant to the topic of this article is that spellcheckers without dynamic compounding mark dynamic compounds as wrong and instead suggest the erroneous split forms. With no access to a spellchecker from the late 1980s, the "corrections" are taken from Google Docs.[1]


UNESCO
I DAG ER DET DEN INTERNASJONALE LÆRER DAGEN
Jørgen G. Bosoni ▸ Astronomer mot orddeling (AMO)
2014 čakčamánnu 26
Billigkroken
Dyre leker

I dag er det den internasjonale lærerdagen. Dette er er en dyreleke.

I dag er det den internasjonale lærer dagen. Dette er er en dyre leke.

Fig. 1: Norwegian compound errors posted in the Facebook group „Astronomer mot orddeling" (Astronomers against split compounds)

The grammatical method became available in the 1990s, for example in Lingsofts spellcheckers for the Nordic languages, and was integrated in Microsoft Word. In this model, there were explicit rules for compounding, and the spellcheckers were thus able to accept nonlexicalized compounds. The problem of dynamic compounding was then solved. Unfortunately, the solution introduced problems with overgeneration, leading to false negatives (unrecognised typos), like the Norwegian common typo in (2), where the correct form would be the adverb *nettopp* "recently, exactly, perfectly", but the typo is disguised by the spellchecker as an absurd compound.

(2) *netopp
ne-topp
old.moon-peak
"the peak of (the lunar phase) old moon"

[1] In fairness it must be added that Google Docs fared better than the spellcheckers of the 1980s in that it was able to recognise the plural form *dyreleker* but it still failed on the singular *dyreleke*.

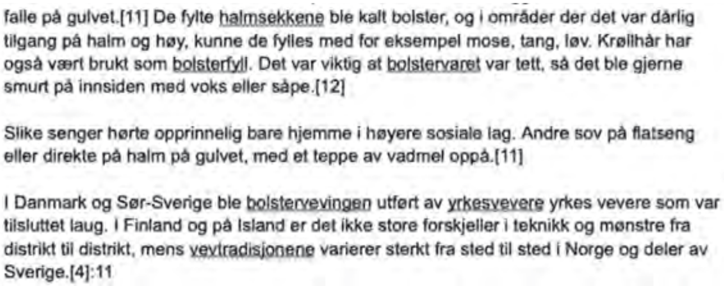
Allowing non-existing compounds of this type into the suggestion mechanism would, of course, add to the problem, since arbitrary compounding of short words in most cases would appear nonsensical and even mislead users into wrong writing habits. The obvious answer to this would be to block dynamic compounding with short words, e.g. 1-3 letter words, but keep it for longer words, like the rare but attested ones in (3):

(3) brettseglingsferie “surfing vacation”
kunnskapstype “knowledge type”
plosivgeminat “plosive geminate”

An even more drastic step would be to block dynamic compounding from the suggestion mechanism altogether.

Instead of efforts aiming at solving these problems, we now unfortunately see a return to spellcheckers based upon attested wordforms only, with Google as its main proponent.

One may think the the solution for word- and text-based approaches is “more text”, and yes, more text does help. The following two figures show text from Wikipedia in Norwegian Bokmål, corrected first by Google Docs and then by giella-nob, a spellchecker based on a finite-state transducer for Norwegian Bokmål.[2] The text contains no typos.



falle på gulvet.[11] De fylte halmsekkene ble kalt bolster, og i områder der det var dårlig tilgang på halm og høy, kunne de fylles med for eksempel mose, tang, løv. Krøllhår har også vært brukt som bolsterfyll. Det var viktig at bolstervaret var tett, så det ble gjerne smurt på innsiden med voks eller såpe.[12]

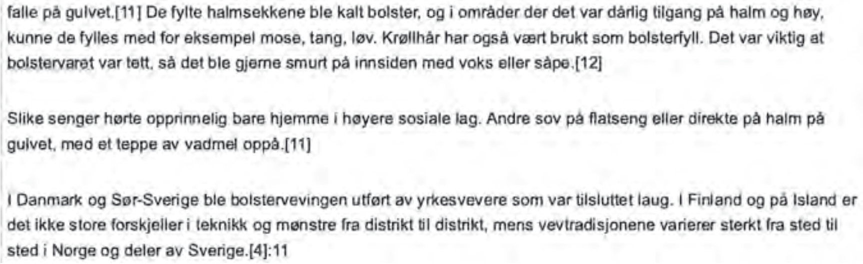
Slike senger hørte opprinnelig bare hjemme i høyere sosiale lag. Andre sov på flatseng eller direkte på halm på gulvet, med et teppe av vadmel oppå.[11]

I Danmark og Sør-Sverige ble bolstervevingen utført av yrkesvevere yrkes vevere som var tilsluttet laug. I Finland og på Island er det ikke store forskjeller i teknikk og mønstre fra distrikt til distrikt, mens vevtradisjonene varierer sterkt fra sted til sted i Norge og deler av Sverige.[4]:11

Fig. 2: Norwegian Bokmål Wikipedia text, corrected by Google Docs

Most of the alleged typos are rare words, linked to traditional handicrafts in pre-industrial times. None of them is found in the 750 million word corpus NoWaC “Norwegian Web as a Corpus” created by the University of Oslo. The spellchecker based on the finite-state transducer allows for dynamic compounding. The false positive *bolstervaret* is due to the noun *var* being blocked from dynamic compounding given that it contains only 3 letters.

[2] https://giellalt.github.io/lang-nob/.



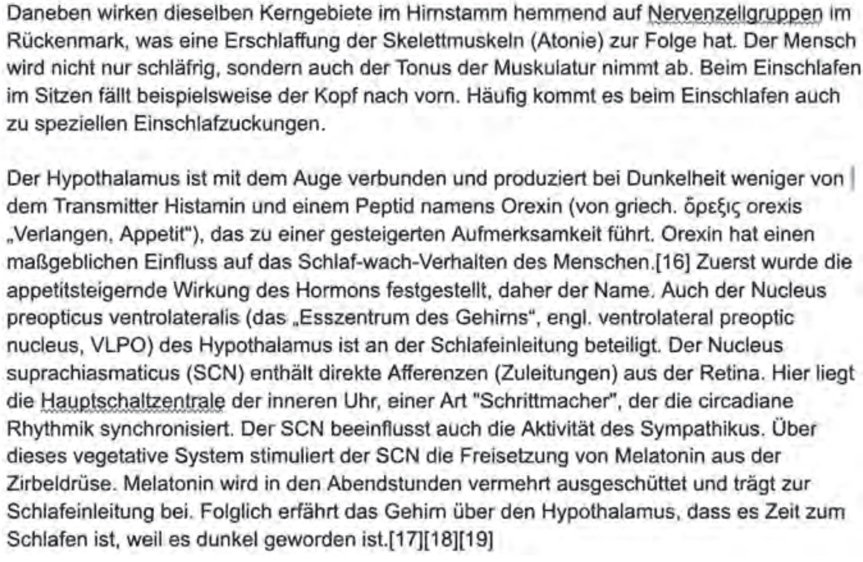
falle på gulvet.[11] De fylte halmsekkene ble kalt bolster, og i områder der det var dårlig tilgang på halm og høy, kunne de fylles med for eksempel mose, tang, løv. Krøllhår har også vært brukt som bolsterfyll. Det var viktig at bolstervaret var tett, så det ble gjerne smurt på innsiden med voks eller såpe.[12]

Slike senger hørte opprinnelig bare hjemme i høyere sosiale lag. Andre sov på flatseng eller direkte på halm på gulvet, med et teppe av vadmel oppå.[11]

I Danmark og Sør-Sverige ble bolstervevingen utført av yrkesvevere som var tilsluttet laug. I Finland og på Island er det ikke store forskjeller i teknikk og mønstre fra distrikt til distrikt, mens vevtradisjonene varierer sterkt fra sted til sted i Norge og deler av Sverige.[4]:11

Fig. 3: Norwegian Bokmål Wikipedia text, corrected by giella-nob

For a national language like Norwegian Bokmål, Google is thus not able to collect enough text to produce a reliable spellchecker. More available text does help, though. Figure 4 gives an example of German scientific text, containing no typos but technical terms, loanwords and even some English and Greek. The latter would, of course, have been out of reach for all but text-based approaches. There are two false positives, though: *Nervenzellgruppen* and *Hauptschaltzentrale*. The two words stand out as being the only 3-part dynamic compounds in the text. Even the resources available for German, the largest language in Europe, is thus not enough to cover words like these.



Daneben wirken dieselben Kerngebiete im Hirnstamm hemmend auf Nervenzellgruppen im Rückenmark, was eine Erschlaffung der Skelettmuskeln (Atonie) zur Folge hat. Der Mensch wird nicht nur schläfrig, sondern auch der Tonus der Muskulatur nimmt ab. Beim Einschlafen im Sitzen fällt beispielsweise der Kopf nach vorn. Häufig kommt es beim Einschlafen auch zu speziellen Einschlafzuckungen.

Der Hypothalamus ist mit dem Auge verbunden und produziert bei Dunkelheit weniger von dem Transmitter Histamin und einem Peptid namens Orexin (von griech. ὄρεξις orexis „Verlangen, Appetit"), das zu einer gesteigerten Aufmerksamkeit führt. Orexin hat einen maßgeblichen Einfluss auf das Schlaf-wach-Verhalten des Menschen.[16] Zuerst wurde die appetitsteigernde Wirkung des Hormons festgestellt, daher der Name. Auch der Nucleus preopticus ventrolateralis (das „Esszentrum des Gehirns", engl. ventrolateral preoptic nucleus, VLPO) des Hypothalamus ist an der Schlafeinleitung beteiligt. Der Nucleus suprachiasmaticus (SCN) enthält direkte Afferenzen (Zuleitungen) aus der Retina. Hier liegt die Hauptschaltzentrale der inneren Uhr, einer Art "Schrittmacher", der die circadiane Rhythmik synchronisiert. Der SCN beeinflusst auch die Aktivität des Sympathikus. Über dieses vegetative System stimuliert der SCN die Freisetzung von Melatonin aus der Zirbeldrüse. Melatonin wird in den Abendstunden vermehrt ausgeschüttet und trägt zur Schlafeinleitung bei. Folglich erfährt das Gehirn über den Hypothalamus, dass es Zeit zum Schlafen ist, weil es dunkel geworden ist.[17][18][19]

Fig. 4: German Wikipedia text corrected by Google Docs

3. The text corpus and the explicit norm

Looking at the problems with the text-based approach in more general terms, the false positives shown here may be seen as an out-of-vocabulary problem. This problem is obviously worse for languages with dynamic compounding than for languages without. Even though it is of no help to the large group of North European languages, at least one may think that a language without compounding and with little morphology would probably get a good spellchecker with far less text than what is available for Norwegian Bokmål.

But the problem is far worse than this. The underlying assumption when basing correction on attested forms is that *the text collection equals the norm*. This implies a principled exclusion of language normative work done by normative bodies, indeed a principled exclusion of language planning as such. The role of normative language institutions is (among many other things) to give advice on how to spell words. The question is thus whether the set of available text collections could be seen as a de facto norm, replacing the explicitly stated norm. Such a move will no doubt result in proofing tools that can help writers "write like all the others", but for normative bodies the answer cannot be but negative.

Proofreaders will tell us that people do make mistakes in writing. Unfortunately, proofreaders are an endangered species. More and more texts are published without proofreading. The democratisation of publishing that came with computers and the internet clearly has its downsides: abolishing typographers has given us ugly typography and abolishing proofreaders has given us more typos. Developing proofing tools from collected texts is thus becoming increasingly problematic. Ideally, the collected texts should, of course, be error free, but this is, to an increasing extent, not the case for publicly available text. Whereas correct forms in most cases outnumber incorrect forms for majority languages (due to fairly good writing skills and huge amounts of text), minority language communities face the double challenge of poorer writing skills and far less text where the correct forms could outweigh the typos.

For minority languages like South Saami, with fewer than 500 speakers, there is another problem. Corpora available for such languages do not even number millions of words. There is also no point in waiting for larger corpora: Small language communities simply do not have enough writers to write the amount of text available for German or Norwegian. Typologically, minority languages often have quite complex morphologies, with a high ratio of words occurring only once in the corpus. For large and more stable written languages, it is to be hoped that the errors would be outnumbered by correct forms, but this is not the case for minority languages.

Furthermore, minority languages typically have young written languages and a norm with a weak status in the language societies concerned. These languages have a marginal position in education and mass media and the normative bodies behind the standards have few ways of enforcing the norm. The key to mastering a written standard is to be exposed to it via extensive reading. Minority languages are predominantly oral, and these languages are rarely used for commercial billboards, film subtitles, etc. The written norm often has a weak status and mother tongue speakers of minority languages tend to choose forms outside the standard. A large percentage of L2 writers also leads to both spelling and grammatical errors. For minority language communities, there is, thus, no way that a collection of texts can set the norm.

4. Tech giants and language communities

Even though the number of languages for which Microsoft and Google offer support is increasing, it is still small: Windows 11 has localisation and proofing for 85 languages and Google Translate is available for 108, when there are 3,514 languages for which there is a translation of at least the New Testament.

Microsoft is making it increasingly harder for third-party providers to add proofing tools to Microsoft Word. With Google, it has always been impossible. The single most important tool for a normative body to implement its norm among writers is the spellchecker. The normative body would thus want to control the content of the spellchecker and it will thus often not be satisfied with the proofing tools offered by the large companies. Moreover, the 3,400 ignored language communities will not get any proofing tools. The result is that the most central common infrastructure of any society, its language, is outside the control of the society to which it belongs.

As language societies, we should not accept being governed by large computer companies. What we need is an independent language technology. The large companies should, of course, make their language tools as they see fit, but they should not prevent language communities from making and distributing their own.

An independent language technology will construct explicit language models. It can take data into consideration but will not be data driven. When needed, the language models will be built as a set of explicit linguistic rules. Such models are transparent: it is possible to correct the models when they make mistakes or when we want them changed due to changes to the language norm. Language corpora are certainly not irrelevant, as any language planner knows. But rather as being seen as The Norm, they should be given the role as a test bench, a reality check: Where should we invest our normativity efforts? What is the balance between linguistic development and language norm? For terminology and vocabulary: *what is actually in use?*

This view has consequences for the relation between language and computers. As language societies, we cannot accept that the very thing that constitutes us as such societies, our language, is beyond our control. Thus, our language models must be made available to the language communities, via the word processing programs that the communities use. The large technology companies have taken it upon themselves to carry the infrastructure of our societies. For this contract to be upheld, they cannot treat language as if it were any commodity. It is not.

An independent language technology can be made in many ways. The main criteria are transparent code and the possibility of governing its properties, thus explicitly deciding the norm.

Our experiences at UiT in Tromsø in Norway are as follows: We work on complex languages with little text, in other words, we work on average human languages. We model the lexicon, compounding, derivation and inflection as finite state transducers. Syntactic analysis and language advice to writers involving sentence or text context is modelled as constraint grammar. This is then integrated in text processing programs (if possible), with good results.

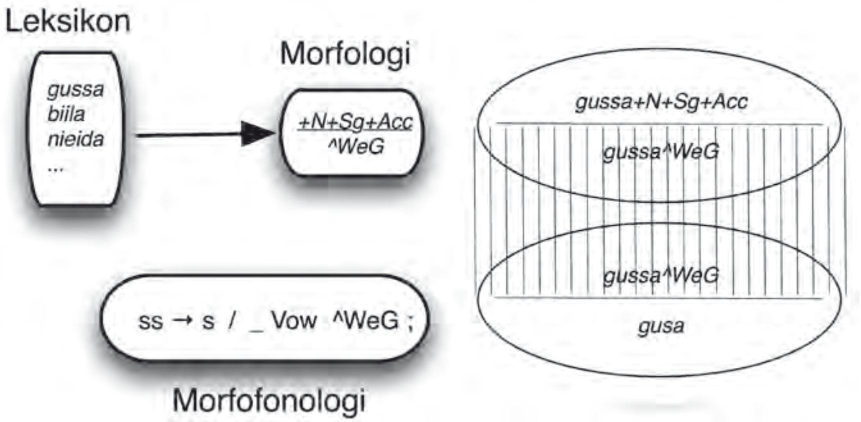
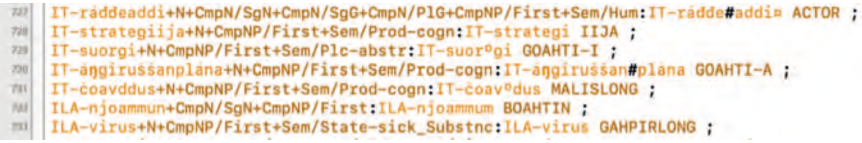


Fig. 5: Finite state transducers as language models for North Saami

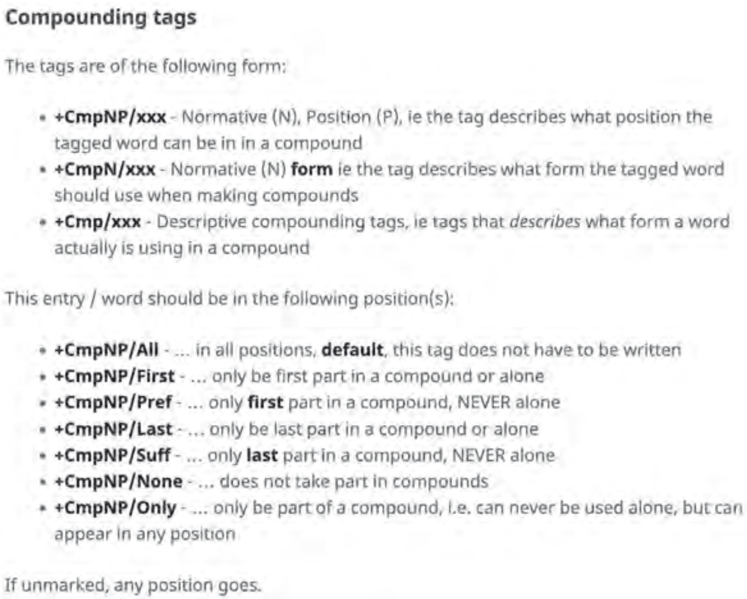
We explicitly govern dynamic compounding by adding tags to the lexicon, as in Figure 6.



```
IT-ráddeaddi+N+CmpN/SgN+CmpN/SgG+CmpN/PlG+CmpNP/First+Sem/Hum:IT-rádde#addi¤ ACTOR ;
IT-strategiija+N+CmpNP/First+Sem/Prod-cogn:IT-strategi IIJA ;
IT-suorgi+N+CmpNP/First+Sem/Plc-abstr:IT-suor°gi GOAHTI-I ;
IT-áŋgirušsanplána+N+CmpNP/First+Sem/Prod-cogn:IT-áŋgirušsan#plána GOAHTI-A ;
IT-čoavddus+N+CmpNP/First+Sem/Prod-cogn:IT-čoav°dus MALISLONG ;
ILA-njoammun+CmpN/SgN+CmpNP/First:ILA-njoammum BOAHTIN ;
ILA-virus+N+CmpNP/First+Sem/State-sick_Substnc:ILA-virus GAHPIRLONG ;
```

Fig. 6: Tags governing compound behaviour, North Saami lexicon

The compound tags are defined in Figure 7.



Compounding tags

The tags are of the following form:

- **+CmpNP/xxx** - Normative (N), Position (P), ie the tag describes what position the tagged word can be in in a compound
- **+CmpN/xxx** - Normative (N) **form** ie the tag describes what form the tagged word should use when making compounds
- **+Cmp/xxx** - Descriptive compounding tags, ie tags that *describes* what form a word actually is using in a compound

This entry / word should be in the following position(s):

- **+CmpNP/All** - ... in all positions, **default**, this tag does not have to be written
- **+CmpNP/First** - ... only be first part in a compound or alone
- **+CmpNP/Pref** - ... only **first** part in a compound, NEVER alone
- **+CmpNP/Last** - ... only be last part in a compound or alone
- **+CmpNP/Suff** - ... only **last** part in a compound, NEVER alone
- **+CmpNP/None** - ... does not take part in compounds
- **+CmpNP/Only** - ... only be part of a compound, i.e. can never be used alone, but can appear in any position

If unmarked, any position goes.

Fig. 7: Compound tags (cf. https://giellalt.github.io/lang-sme/src-fst-root.lexc.html)

Others may do it differently. This is fine, as long as your language model does what you want, and you are able to put it into use in the word processor. What we do at UiT is openly available for adaption and reuse at https://giellalt.github.io/.

5. Conclusion

Normative language work must be independent from and stand above actual language use. This calls for an explicit and transparent language technology. Such a language technology is threatened from two sides: from the dominant trend within AI, favouring data-driven approaches, and from the major programming houses, preventing third-party language technology programs from being integrated in their word processor software. As shown here, an alternative path is possible: to develop transparent open source rule-based systems that can be easily integrated into the linguistic software of the big tech companies. The issue is too important to let slip.

Andreas Witt/Paweł Kamocki

The future is now. The digital transformation in the German linguistics community and the key role of the IDS

Abstract

The Leibniz-Institute for the German Language (IDS) was established in Mannheim in 1964. Since then, it has been at the forefront of innovation in German linguistics as a hub for digital language data. This chapter presents various lessons learnt from over five decades of work by the IDS, ranging from the importance of sustainability, through its strong technical base and FAIR principles, to the IDS' role in national and international cooperation projects and its expertise on legal and ethical issues related to language resources and language technology.

1. Introduction

The Leibniz Institute for the German Language (Leibniz-Institut für Deutsche Sprache, hereinafter: IDS) is the central academic institution for the study and documentation of the contemporary usage and recent history of the German language.[1] Since its establishment in 1964, the IDS has been at the forefront of innovation in German linguistics.

This chapter describes the IDS' role as a digital hub for German language data (Section 1) and presents several "lessons learnt" from the (nearly) sixty years of its existence (Sections 2–7). These sections focus on the importance of digital language data for the IDS (Section 2), the importance of sustainability in its many dimensions (Section 3), and the role of the technological base (Section 4). The remaining sections present the efforts made at the IDS towards recognising and addressing the user's specific needs when it comes to language data and tools (Section 5), and making the data and tools findable– a task considerably facilitated by international and national cooperation projects (Section 6), where legal and ethical issues are one of the focal points (Section 7).

2. The IDS as a digital hub

The IDS was established in 1964 in the city of Mannheim. The choice of this city was not accidental, as Mannheim has long had strong links with German linguistics.

[1] Cf. https://www.ids-mannheim.de/?id=1491&L=1 (last accessed 04-05-2022).

This is where the seat of the *Bibliographisches Institut*, publisher of the Duden dictionary, was moved in 1953, together with the *Institut*'s large archive. Although, after the reunification of Germany, the main seat of the publishing house was relocated to Berlin, its Language Technology Division remained in Mannheim. In 2013, the *Bibliographisches Institut*'s archive was donated to the library of the University of Mannheim. Moreover, Mannheim is also the city where the Council of German Orthography[2] (established in 2004) is based.

The IDS was not only established in a very special place but also in a very special era. In the 1960s, Germany's state propaganda was still in living memory, and the IDS was committed from the start to strict empiricism, which was a politically innovative approach at the time. In this spirit, the IDS follows a descriptive rather than a prescriptive approach to language research. The choice of digital methods, with their cold objectivism, is one of the ways to guarantee freedom from any ideological influences.

It is therefore only natural that the IDS – home to the so-called Mannheim School of Corpus Linguistics (Teubert/Belica 2014) – has long been at the forefront of the digital transformation of language research in Germany. Shortly after its establishment, in the very early days of corpus linguistics, the IDS began collecting German texts in digital form, initially using punch cards as data carriers. The first electronic corpus of German, the *Mannheimer Korpus I* (MK I), completed in 1969, was compiled in this way; it numbered 2.2 million words in 293 texts. Another corpus, LIMAS (*Linguistik und Maschinelle Sprachbearbeitung*) was compiled between 1970 and 1971; it consisted of 500 texts divided into 33 subject areas. In 1975, this and other early IDS corpora were printed on continuous form paper; they are stored to this day in this form in the IDS archive (Fürbacher et al. 2017).

This long tradition of text corpora at the IDS led to the creation of DeReKo (*Das Deutsche Referenzkorpus*), the world's largest collection of German texts designed for language research (Kupietz et al. 2018). As of March 2022, DeReKo contains 53 billion words – and is growing at a steady pace. DeReKo is available for online querying by registered users (the registration process is free and simple) via COSMAS II (Corpus Search, Management and Analysis System)[3] and KorAP (Corpus Analysis Platform).[4] DeReKo is also subdivided into smaller sub-corpora according to various criteria, thereby catering to the users' specific needs (cf. Section 5 below). It has been an inspiration for numerous other national language reference corpora in Europe.

Not only text data but also speech data have been collected at the IDS. In 1971, the German Speech Archive (*Deutsche Spracharchiv*, DSAv, compiled since 1932)

[2] http://www.rechtschreibrat.com (last accessed 01-07-2022).

[3] https://cosmas2.ids-mannheim.de/cosmas2-web/ (last accessed 01-07-2022).

[4] https://korap.ids-mannheim.de (last accessed 01-07-2022).

was transferred to the IDS. Later, it became the Archive of Spoken German (*Archiv für Gesprochenes Deutsch*),[5] which is still being added to today (Fürbacher et al. 2017).

The directors of the IDS and their affinity for digital matters played a crucial role in the institution's becoming a hub for digital language data. Before he was appointed director of the IDS (a position he occupied between 1976 and 2002), Prof. Dr. Gerhard Stickel worked as a researcher at the German Computing Centre (*Deutsche Rechenzentrum*, DRZ) in Darmstadt; he was also involved in early-stage AI research. Prof. Dr. Ludwig M. Eichinger (director of the IDS between 2002 and 2018) already used the IDS' digital data as a PhD Student. Prof. Dr. Henning Lobin (Director of the IDS since 2018) obtained his habilitation in Computational Linguistics at the University of Bielefeld in 1996, and then served as Professor of Applied and Computational Linguistics at the Justus Liebig University in Giessen for nearly two decades. This proves that since the institution's early days, the IDS' directors understood the importance of digital technology and realised its potential for language research. Some distinguished members of the IDS' Scientific Advisory Board, like Hans Uszkoreit and John Nerbonne, were also pushing the institution up the digital path.

In 2019, a dedicated Department for Digital Linguistics (*Digitale Sprachwissenschaft*) was created at the IDS, headed by one of the co-authors of this chapter. As of mid-2022, there are eighteen researchers in the department, working on the collection and curation of language data, the long-term archiving of language data, and national and international infrastructure projects as well as legal and ethical issues related to the above-mentioned domains (cf. Section 7). The establishment of the department was crucial for infrastructure projects at the IDS. The department's associates are (or were) involved in such projects as D-SPIN[6] (the predecessor of CLARIN-D,[7] the German national branch of CLARIN ERIC (see below), and later CLARIAH-DE[8]), TextGRID (a virtual research environment for the humanities optimised to work with TEI-coded resources),[9] *Verwertug Geist*[10] (exploring the potential of knowledge transfer in the humanities and related domains) and Text Transfer[11] (on the application of corpus-based methods to predict the impact of scientific texts).

[5] https://agd.ids-mannheim.de/index.shtml (last accessed 01-07-2022).

[6] https://weblicht.sfs.uni-tuebingen.de/publikationen.shtml (last accessed 06-07-2022).

[7] https://www.clarin-d.net/en/ (last accessed 06-07-2022).

[8] https://www.clariah.de/en/ (last accessed 06-07-2022).

[9] https://textgrid.de/en/ (last accessed 06-07-2022).

[10] https://www.ids-mannheim.de/fi/abgeschlosseneprojekte/verwertung-geist/ (last accessed 06-07-2022).

[11] https://www.ids-mannheim.de/fi/projekte/texttransfer/ (last accessed 06-07-2022).

3. The role of sustainability

The many resources, tools and activities mentioned in the previous section could not have been developed at the IDS if the institution had not provided sufficient guarantees of sustainability.

Organizational sustainability is a pre-condition of trust. It is indeed hard to trust an organization that cannot guarantee its survival over a long period of time. This is clearly visible in the world of education, where older establishments (such as Oxford and Cambridge universities, among the first universities in the world) have an obvious reputational advantage over newly created ones, no matter how generously funded and how enthusiastically advertised they are. The fact that an establishment has been issuing internationally recognized diplomas for decades if not centuries is perceived as a guarantee that a diploma from this establishment will retain its value in the foreseeable future. The same reasoning also applies to research organisations.

According to the Practical Guide for Sustainable Research Data recently published by Science Europe (2021),[12] the sustainability of a Research Performing Organisation (RPO) is to be evaluated in the following areas: Organisational Engagement and Commitment; Policy Environment; Financial Aspects; Training; Technical Preparedness; and Communication and Awareness Raising. As a research institution with over 50 years of tradition and stable sources of funding (the German Federation and the Federal State of Baden-Wuerttemberg), the IDS has a high score in all of the above-mentioned domains.

In a narrower sense, sustainability refers specifically to the perennial archiving of research data (technical sustainability), i.e., providing guarantees that the data will be re-usable and available (preferably at their original location, even if the data themselves are marked as outdated) over long periods of time. This is achieved via the standardization of data formats, and especially via continuous conservative (or, rather, preservative) development. Both of these necessitate a strong technological base.

4. The role of a strong technological base

Another lesson learnt from the IDS' experience as a digital hub for language data concerns the importance of a strong technological base.

The implementation of the current technological base at the IDS has been described by Witt/Schonefeld (2011). The authors identify the following aspects of the technological base:

[12] https://scienceeurope.org/media/b3odxx3s/se-practical-guide-sustainable-research-data.pdf (last accessed 01-07-2022).

- *Services* provided to users are the most important part of the infrastructure; they include internet access (e.g., via Eduroam), e-mail, cloud storage, virtual workspace (e.g., an online text editor), an online library catalogue, etc. During the COVID-19 pandemic and generalized home office, it has become particularly important for services to be available not only on site, but also remotely, via virtual private networks (VPN);
- *Identity Management*: access to most services requires user authentification; it is simplified considerably if the user's personal data are managed centrally (e.g., by the HR department), and each service synchronises its access data with a central identity database. This minimizes the risk of errors due to typos, facilitates the recommended periodic changes of passwords and changes of usernames (e.g., following marriage), and enables the accounts of former employees to be deleted quickly;
- *Operating and Maintenance*: all components of the technological base (servers, workstations, internet connection, printers, etc.) should be classified according to their importance; critical components (such as the internet connection) should be backed up by redundant systems, and the whole infrastructure should be constantly monitored so that immediate action can be taken in case of failure;
- *Security*: research data, especially those held by a language research institution, may be thought of as presenting little to no interest for hackers; this, however, is not true. Many attacks are quantity-oriented and their perpetrators simply want to affect as many computers as possible, regardless of their "quality". Moreover, IT security is increasingly a legal requirement for storing and processing personal data (cf. Articles 5.1(f) and 32 of the General Data Protection Regulation) and corpora based on the Text and Data Mining exception (Article 3.2 of the 2019 Directive on Copyright in the Digital Single Market). Protection against unauthorized access is, therefore, an essential feature of a strong technological base in a research institution.

5. Transfer depends on technology

Language institutes, just like any other establishments or organisations, should never lose sight of the needs of their "customers" (no matter whether they are called "clients", "users" or "target groups", the idea remains the same).

In the case of language institutes, this is particularly difficult as the "customers" are indeed particularly difficult to define. It might be tempting to say that a language institute's work is carried out "for science", "for the greater good" and "for future generations" – all of these are true – but there are also actual people, here and now, who can benefit from the results of a language institute's work.

The first and largest group of the IDS' customers is undoubtedly other research institutions, and especially universities: the place where future teachers of national languages are educated. Public administrations also count among the IDS' "clients". On occasions, the IDS also works with the private sector, such as the publishers of dictionaries and encyclopaedias, interested in keeping their publications up to date. This pool of "clients" is expected to grow steadily, as more and more actors realize that the ability to process and analyse digital text data is an important component of 'digital literacy', a fundamental skill in the contemporary world and not just limited to the job market.[13]

Each of the above-mentioned groups has its specific needs which the IDS is trying to cater for. In particular, representatives of each of these groups expect to receive empirical data pre-processed in a specific way. Responding to this expectation requires skilful usage of the possibilities offered by digital technology.

6. The role of national and international cooperation

There is a great variety of language data and language corpora. They can be divided according to their modality (text, speech, audio-visual data), their context (e.g. parliamentary debates, poetry, everyday speech, simplified language, L2 and learner's speech), their time periods and their media (e.g., computer-mediated communication). This variety makes it complicated for users to find the exact type of resource that corresponds to their needs. In order to facilitate this task, it is crucial for the resources to be marked with appropriate metadata.

However, even a very complete metadata description of a language resource does not guarantee that it will be found and re-used. The metadata should also be 'advertised' through appropriate channels, such as catalogues provided as part of national and international cooperation projects.

National and international cooperation in the field of language resources is, to a large extent, motivated by the idea of FAIR data, i.e. making research data Findable, Accessible, Interoperable and Re-Usable (Wilkinson et al. 2016). Cooperation between institutions, within and across borders, is necessary to achieve this ideal, as it allows them to mutualize and coordinate efforts towards addressing some of the common problems, such as legal and ethical issues (cf. Section 7).

The IDS has been involved in CLARIN (Common Language Resources and Technology Infrastructure, formally established in 2012[14]) since its conception phase. CLARIN's mission is to create an online environment in which digital language resources and tools from all over Europe are accessible through a single sign-on for researchers in the humanities and social sciences (Fišer/Witt 2022).

[13] Cf. the podcast by Andreas Witt and Thorsten Meyer as part of the Max Planck Society's series 'Digital Qualifiziert', recorded in 2021, available at: https://soundcloud.com/max-planckgesellschaft/digital-qualifiziert-andreas-witt-thorsten-meyer (last accessed 06-07-2022).

[14] By the Commission Decision 2012/136/EU of 29 February 2012 setting up the Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium (CLARIN ERIC).

The IDS is a certified CLARIN B-Centre providing long-term storage of Germanic language resources. Apart from the storage facilities, the IDS' contribution to CLARIN involves the institution's expertise in language archives, linguistic tools, long-term preservation, multimedia and multimodal data as well as legal and ethical issues.

The IDS also plays an important role in the Text+ Consortium, whose goal it is to preserve text- and language-based research data in the long term and enable their broad use in science.[15] Formally established in 2021, Text+ has been approved as a consortium for the nationwide initiative to create a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI),[16] based on an application submitted by the applicant institution, the IDS, and the four co-applicant institutions, the Berlin-Brandenburg Academy of Sciences and Humanities, the German National Library, Göttingen State and University Library, and the North Rhine-Westphalian Academy of Sciences, Humanities and the Arts. Apart from the five applicants, more than 25 additional participating institutions contribute their specialist expertise to the initiative, a number which is expected to grow. Erhard Hinrichs – who, alongside his full professorship for General and Computational Linguistics at the University of Tübingen, is also affiliated to the IDS – serves as the spokesperson for the Text+ consortium.

The IDS is also part of many smaller international infrastructure projects, such as DeutUng (Deutsch-ungarischer Sprachvergleich) (with the University of Szeged, Hungary)[17] and DRuKoLA (Deutsch-Rumänische korpuslinguistische Analyse) (with the University of Bucharest and the research institutes of the Romanian Academy in Bucharest and Iaşi)[18] (Kupietz et al. 2019a; Cosma et al. 2016). Both these projects are integrated in a larger EuReCo (The European Reference Corpus) (launched in 2012), which aims to virtually join various national reference corpora by using the same analysis platform, KorAP (cf. above) (Kupietz et al. 2019b; Trawiński/Kupietz 2021).

7. The importance of legal and ethical issues

Since its establishment, the IDS has handled third-party language data, especially provided by such entities as the press and book publishers (cf. Section 2). Re-use of such data for research purposes requires a careful assessment of its legal status. Thanks to the experience acquired over the decades, the IDS has become a national (and, to a certain extent, a European) centre of expertise on the many legal and ethical issues affecting language resources.

[15] https://www.text-plus.org/en/home/ (last accessed 07-07-2022).

[16] https://www.nfdi.de/?lang=en (last accessed 07-07-2022).

[17] https://www.ids-mannheim.de/gra/projekte/deutung/ (last accessed 06-07-2022).

[18] https://www.ids-mannheim.de/digspra/kl/projekte/drukola (last accessed 06-07-2022).

As a general rule, language data are protected by copyright, as language expressions are in fact the result of their authors' own intellectual creations.[19] Copyright protection expires 70 years after the death of the author, so in principle all born-digital language data are still in copyright. The re-use (reproduction of and communication to the public) of such data requires permission from the copyright holder (i.e., typically, the author, their descendents or the publisher, if copyright was transferred by the author), unless it is expressly allowed by a statutory exception. Such exceptions exist and they are currently expanding (e.g., new exceptions for Text and Data Mining purposes were introduced by the 2019 Directive on Copyright in the Digital Single Market) but they are accompanied by complex requirements which, according to the principle *exceptio est strictissimae interpretatonis*, always need to be interpreted narrowly. This means that before an exception can be relied on, a thorough analysis of each specific case is necessary.

When copyright exceptions are insufficient for the intended use, it is necessary to negotiate a license (Latin: permission) with the copyright holders, which also needs to be carefully drafted and interpreted. On the other hand, researchers who want to make data and content generated by them (e.g., research articles, software tools) available for re-use by granting up-front permission to every member of the public, in the spirit of the Open Access/Open Data/Open Science movements, can achieve this via proper licensing, using the so-called public licenses, such as Creative Commons (CC) or the General Public License (GPL). The use of such licenses for research results is increasingly required by research funding bodies.

Moreover, language data often contain personal data, i.e., as per the legal definition (Article 4, (1) of the GDPR), "*any information relating to an identified or identifiable natural person*". The processing of such data, even for research purposes, must abide by the strict framework of the GDPR, which affects especially speech and multimodal resources.

All these issues need to be properly addressed already in the conception phase of a language research project. For this reason, it is important to not only provide researchers with guidance and advice but also to educate them so that they are able to identify potential friction points at an early stage. Therefore, the IDS' legal experts not only provide Legal Helpdesk services for CLARIN but also created the Legal Information Platform.[20] Moreover, they have been involved in the creation of two LegalTech tools destined specifically for researchers in the data-intensive humanities and social sciences: the Public License Selector[21] (Kamocki et al. 2016) and the Consent Form Wizard.[22] Recently, the IDS has also published a set of hand-

[19] Cf. CJEU's judgement of 16 July 2009 in the case C-5/08 (Infopaq).

[20] https://www.clarin.eu/content/legal-information-platform (last accessed 01-07-2022).

[21] https://github.com/ufal/public-license-selector (last accessed 01-07-2022).

[22] https://consent.dariah.eu/ (last accessed 01-07-2022).

outs on GDPR compliance, specifically addressing issues related to language research and the archiving of language resources.[23]

Finally, ethical issues are also of growing importance for language resources and language technology. Despite a growing number of ethics-related concerns, the exact content of ethical principles governing language resources and language technology remains unclear. In order to mitigate this, the IDS researchers and authors of this chapter have proposed a tentative taxonomy of ethical issues in the sector, based on five principles: Privacy, Property, Equality, Transparency and Freedom (Kamocki/Witt 2022).

8. Conclusion

As demonstrated above, the IDS plays a key role in digital transformation in the German language community. With its stable and sustainable funding, over half a century of experience with collecting, curating and archiving language data, a rich and diversified portfolio of projects and activities, strong participation in international initiatives and, last but not least, a dedicated department of Digital Linguistics, the IDS is well equipped to assume this role for decades to come.

References

Cosma, R./Cristea, D./Kupietz, M./Tufiş, D./Witt, A. (2016): DRuKoLA – Towards contrastive German-Romanian research based on comparable corpora. In: *4th Workshop on Challenges in the Management of Large Corpora. Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 28–32.

Fišer, D./Witt, A. (eds.) (2022): *CLARIN: The infrastructure for language resources*. (= Digital Linguistics 1). Berlin: de Gruyter.

Fürbacher, M./Varadi, T./Witt, A. (2017): Digitale Forschungsinfrastrukturen: ihre Nutzung durch die Mitglieder der Europäischen Föderation Nationaler Sprachinstitutionen. In: Dąbrowska-Burkhardt, J./Eichinger, L. M./Itakura, U. (eds.): *Deutsch: lokal – regional – global*. (= Studien zur Deutschen Sprache 77). Tübingen: Narr, 103–113.

Kamocki, P./Witt, A. (2022): Ethical issues in language resources and language technology – tentative taxonomy. In: *Proceedings of the 13th Conference on Language Resources and Evaluation* (LREC 2022), Marseille, France. Paris: European Language Resources Association (ELRA), 559–563.

Kamocki, P./Stranak, P./Sedlak, M. (2016): The public license selector: making open licensing easier. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia. Paris: European Language Resources Association (ELRA), 2533–2538.

[23] https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/10695 (last accessed 01-07-2022).

Kupietz, M./Cosma, R./Witt, A. (2019a): The DRuKoLA project. In: Cosma, R./Kupietz, M. (eds.): *On design, creation and use of the Reference Corpus of Contemporary Romanian and its analysis tools. CoRoLa, KorAP, DRuKoLA and EuReCo*. Bucharest: Editura Academiei Române.

Kupietz, M./Margaretha, E./Diewald, N./Lüngen, H./Frankhauser, P. (2019b): What's new in EuReCo? Interoperability, comparable corpora, licensing. In: *Proceedings of the Workshop on Challenges in the Management of Large Corpora* (CMLC-7, 2019), Cardiff, UK. Mannheim: Institut für Deutsche Sprache, 33–39.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: new developments – new opportunities. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan. Paris: European Language Resources Association (ELRA), 4354–4360.

Teubert, W./Belica, C. (2014): Von der linguistischen Datenverarbeitung am IDS zur "Mannheimer Schule der Korpuslinguistik". In: *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache.

Trawiński, B./Kupietz, M. (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, H./Witt, A./Wöllstein, A. (eds.): *Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. Jahrbuch des Instituts für Deutsche Sprache 2020*. Berlin/Boston, de Gruyter.

Wilkinson, M./Dumontier, M./Aalbersberg, I. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: *Scientifix Data* 3, No. 160018. https://doi.org/10.1038/sdata.2016.18.

Witt, A./Schonefeld, O. (2011): Informationsinfrastrukturen am Institut für Deutsche Sprache. In: Stickel, G./Varadi, T. (eds.): *Language, languages and new technologies: ICT in the service of languages. Contributions to the Annual Conference 2010 of EFNIL in Thessaloniki*. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 87). Frankfurt a. M. et al.: Lang, 197–211.

Frieda Steurs

The role of the Dutch Language Institute (INT) in the digital age

Abstract

The *Instituut voor de Nederlandse Taal* (or Dutch Language Institute) is the place for anyone who wants to know anything about Dutch through the centuries. The institute collects new Dutch words, updates important reference works such as the *Algemene Nederlandse Spraakkunst*, the main standard work on Dutch grammar, and creates terminology lists to make professional jargon accessible. The institute also takes a central position in the Dutch-speaking world (the Netherlands, Flanders, Suriname and the Netherlands Antilles) as a developer, keeper and distributor of corpora, lexica, dictionaries and grammars. With these sustainable language resources, all the result of scholarly methods, the Dutch Language Institute provides the necessary building blocks for the study of Dutch. In this presentation, we will focus on the structure and development of the central digital language infrastructure and plans for the near future to improve our processes using the most recent insights into computational and corpus-driven linguistics and AI.

1. The Dutch Language Institute: a treasury of Dutch language materials

In 2016, the Institute for Dutch Lexicology was turned into the more broadly oriented Dutch Language Institute (INT). This change went hand in hand with the renewed terms of reference of the General Secretariat of the Dutch Language Union, which was to focus on policy tasks, leaving the executive tasks to the INT. For the Dutch Language Institute, this transfer of tasks provided the opportunity to broaden its own activities. The institute became the central point of contact regarding the keeping and maintenance of digital language materials and the safekeeping of data collections related to any variations of Dutch. This evolution reflects the strongly altered landscape of linguistic research: large language infrastructures are digitally set up and contain corpora, dictionaries and other specialised lexicons and databases, grammar and so much more. The institute develops and provides data for dictionaries, (computational) lexicons, corpora and tools. The dictionaries are accessible online. Software and computational linguistic tools are available open source.

The INT has a central position in the whole of the Dutch-speaking world (the Netherlands, Flanders, Suriname and the Caribbean) as a developer, keeper and distributor of scholarly and sustainable language resources. The institute is well equipped for this task having a large international network for the exchange of information with like-minded institutions. The Dutch Language Institute also provides the necessary building blocks for all language applications aimed at the development and improvement of businesses and public organisations. We intend to strengthen this role in the coming years, which is why we are focusing on the sustainable distribution of any language materials, with an emphasis on:

1) Dutch vocabulary, both historical and contemporary, both in standard language and dialects, both in general language and professional language;
2) new technologies and techniques to make the internet accessible for linguistic research and for the ongoing maintenance of constantly updated, extensive corpora of contemporary Dutch;
3) a contribution to the accessibility of historical text material (coming from inside and outside the INT), in which considerable variations in spelling are no longer a search impediment and ways are offered to detect and circumvent variations in word use;
4) the use of and contribution to new computational linguistic or language technology techniques to help information retrieval from language materials;
5) the formal structuring of linguistic information, making it suitable for computational linguistic applications;
6) a further expansion of spelling information;
7) the realisation of facilities for third parties to contribute interactively to the description of the Dutch language and the optimisation of the central digital data infrastructure for this purpose;
8) becoming a point of contact for all language teachers and building an infrastructure of language materials that are useful and necessary support for teaching Dutch to various types of language learners.

2. The INT in the digital age: CLARIN services

The institute has responded to new developments in the humanities, especially in the field of digital humanities. In order to fulfill this role, the INT maintains a digital infrastructure for Dutch, paying attention to language variation (terminology, dialects, etc.). Both academic and non-academic parties can make use of this infrastructure. The INT sees a clear overlap between its own activities – the central data infrastructure – and recent developments within the e-humanities. With its own expertise, the INT contributes to the digital future of the humanities in the Netherlands and Flanders. On the one hand, knowledge and products are delivered which support other scientific organisations, and on the other hand collaboration with the e-humanities enhances the quality of the central data infrastructure for Dutch. In the next few years we will work closely together with centres for digital humanities at various universities and with networks such as the KNAW Humanities Cluster (Netherlands), Digital Humanities Benelux, and the WOG Digital Humanities (Flanders). The INT functions as a CLARIN[1] centre for Flanders and informs Flemish researchers about the latest developments in the field of linguistic sources and the wider linguistic infrastructure in Europe (CLARIN ERIC).[2] This allows any researcher to learn more about access to repositories, standards, metadata, available corpora, methods to encode their own corpus material, and storage facilities, etc. Researchers and students affiliated with universities and other research institutes can log in with single sign-on (SSO) to use tools and materials. These can be found through portals. This also makes it easy to keep track of ongoing and previously conducted research, which stimulates the cultivation of (international) contacts with fellow researchers.

The portals enable the online use or downloading of tools and data. Researchers have the option of using a personal workspace. Moreover, they can safely and sustainably leave their own research data and research tools in the infrastructure upon finishing their project. Crucially, CLARIN guarantees that tools will be updated and that materials will remain available and researchable through the use of persistent identifiers.

In 2021, we became a CLARIN K-Centre, the K standing for knowledge, focused on Dutch. In this role, the INT also shares its knowledge with non-Dutch researchers.[3] We provide extensive information about Dutch: linguistic properties, language advice, available tools and resources, etymology, and dialects, etc.

Also in 2021, we succeeded in having Belgium join the CLARIN resource network. The Belgian CLARIN consortium CLARIN-BE is led by the INT. Dr Vincent Vandeghinste, senior staff member of the INT, is the national coordinator for CLARIN-BE.

3. CLARIN + DARIAH = CLARIAH

CLARIAH[4] is a large research project in the Netherlands funded by the National Science Foundation. Researchers in the humanities joined forces and combined CLARIN with DARIAH[5] research groups and funding. CLARIAH develops, facilitates, and stimulates the use of digital humanities resources and infrastructures. We offer these resources to researchers and other professionals in an insightful and user-friendly way.

[1] CLARIN stands for Common Language Resources and Technology Infrastructure.

[2] https://www.clarin.eu.

[3] https://kdutch.ivdnt.org/wiki/K-Dutch.

[4] https://www.clariah.nl/.

[5] DARIAH stands for Digital Research Infrastructure for the Arts and Humanities.

This includes tools, particularly software applications and services aimed at digitising, annotating, analysing, and reporting research data. These tools can help researchers to:

- ✓ Perform research tasks faster, more efficiently, and more accurately;
- ✓ Search, edit, analyse, and present large amounts of data;
- ✓ Pose research questions that could not be answered before, for new scholarly insights.

Not only tools but also data sets are made available: these data sets range from handwritten seventeenth-century texts to radio and television recordings as well as social media reports on current developments. They also contain databases with structured data on historical economic parameters, linguistic phenomena, people, and locations, etc.

The work packages in CLARIAH are well distributed across different scientific disciplines and specialisms to develop its digital resources. There are teams with work packages for linguistics, socio-economic history, media studies, textual sources, and (shared) technology.

Some examples of CLARIAH projects[6] are:

NAMES: Dutch corpus of person name variants

Spelling variations, variants, and digitisation errors in person names are serious obstacles for search operations in historical documents. The NAMES project aimed to standardise 564,000 different surnames and 190,113 different given names with the help of the CLARIAH tool TICCL.

NEWSGAC: News Genres Transparent Automatic Genre Classification

How genres in newspapers and television news can be detected automatically using machine learning in a transparent manner to capture the shift from opinion-based to fact-centred reporting.

Bridging the Gap: Digital Humanities and the Arabic-Islamic Corpus

This project harnesses state-of-the-art digital humanities approaches and technologies to make pioneering forays into the vast corpus of digitised Arabic texts. This is primarily done along the lines of two case studies: Islamic jurisprudence and Arabic literature on proselytism.

[6] https://www.clariah.nl/projects.

CLARIAH Flanders

Being a Dutch-Flemish institute, the INT also participates in the CLARIAH Flanders research project funded by the Flemish Research Foundation. CLARIAH-VL is the Flemish contribution to the European research infrastructures DARIAH and CLARIN. Through its partner institutions, CLARIAH-VL helps organise a series of training events such as workshops, summer schools, and lectures. To support the free exchange of knowledge, CLARIAH-VL encourages its members and presenters to make any teaching or training events available to the general public by publishing them under open licenses and sharing them with the community (whenever they are legally allowed to do so).

4. Inclusion and diversity in the digital age

The Dutch Language Institute focuses on developing materials for the Deaf community and for language users with limited literary skills.

4.1 Working for the Deaf community: SignOn – Sign Language Translation Mobile Application and Open Communications Framework7

People who are deaf or hard of hearing face the challenge of interacting with others in real-life situations and are often excluded from accessing information in society. The EU-funded SignON project aims to develop a mobile application that will translate between different European sign and verbal languages. The application, lightweight software running on a standard mobile device, will interact with a cloud-based distributed framework dedicated to computationally heavy tasks. The application and framework will be designed through a co-creation approach where users will work together with the SignON researchers and engineers. The application will be easily adaptable to other languages (sign and spoken) and modalities and will ultimately promote equitable exchange of information among all European citizens.

A large part of the consortium consists of Dutch and Flemish partners, and both the Flemish and Dutch sign language and Dutch play a major part in this project.

4.2 Low literacy and language learners

The Dutch Language Institute has a corpus with data from two newspapers written especially for language learners and people with low literacy: the Wablieft news-

[7] https://cordis.europa.eu/project/id/101017255.

paper[8] (Flanders) and WAI-NOT newspaper[9] (the Netherlands). We use these materials to create new applications for language learning.

At the same time, we cooperate with Oefenen.nl[10], an online environment where adults can practise to improve their basic skills and knowledge. By creating appropriate language materials, we help them study at their own pace.

5. New developments in lexicographic insights: insights in the development of AI for NLP

We participate in the Netherlands AI Coalition (NL AIC), a public-private partnership in which the government, the business sector, educational and research institutions as well as civil society organisations collaborate to accelerate and connect AI developments and initiatives. The ambition is to position the Netherlands at the forefront of knowledge and applications of AI for prosperity and well-being. We continually do so with due observance of both Dutch and European standards and values. The NL AIC functions as the catalyst for AI applications in our country. In 2020, a **workshop on AI for Innovation** was organised by the ministries of both the Netherlands (OCW) and Flanders (EWI) .

The topics covered were:

- AI applied within research in particular on *natural language processing*;
- Smart Industry (Digital Innovation Hubs to introduce AI to companies and public services);
- AI & Legislation (Human-Centric AI);
- Data Sharing (structures and solutions for data sharing);

The Dutch Language Institute provided the input for the first action point.

6. Conclusion

Because current developments in the domains of computational linguistics, NLP, and AI are important to the Dutch Language Institute, it participates in new projects and workshops and implements these new technologies in its work on the digital language infrastructure.

[8] http://www.wablieft.be/nl/krant.

[9] https://www.wai-not.be/page/10.

[10] https://oefenen.nl/.

Marek Łaziński

Polish language resources 2021[1]

Abstract

This paper presents digital resources and language technology in Polish. The Polish LT landscape comprises the National Corpus of Polish with 1.5 billion words, a monitor corpus Monco with 7.7 billion words, several parallel corpora including Polish texts, the Polish WordNet with 600 thousand lexical relations, tools for building and maintaining corpora, taggers, lemmatizers and dependency parsers.

Digital language resources and language technology in Poland

Polish has been present on the web for years. In 2020, the number of .pl domains reached almost 2.5 million, in 2021 the number of internet users in Poland added up to 28.8 million, i.e., 87% of the population. In 2022 Polish Wikipedia ranked 11th in terms of the number of articles (currently over 1.5 million). In 2020, 77 percent of Poles used Facebook and 60 percent were Messenger users.

Since the 1990s several written corpora of contemporary Polish have been created, starting with the National Corpus of Polish: nkjp.pl. Constructed in 2007–2011 by the Institute of Computer Sciences Polish Academy of Sciences, the Institute of Polish Language PAS, the University of Łódź, and the Polish Scientific Publishers PWN, the corpus comprises over 1.5 billion words, with 250 million in the balanced part covering texts from 1918 to 2010. All texts are annotated morphosyntactically, 1 million words in a sub-corpus have been fully annotated manually. Two search programs give access to sophisticated morphosyntactic concordance queries and to a collocations search (Przepiórkowski et al. 2012). The continuation of the National Corpus of Polish is the Corpus of the Decade, a project in progress (http://korpus-dekady.ipipan.waw.pl).

There are many parallel corpora with Polish: Polish-English (http://paralela.clarin-pl.eu), Polish-German (http://diaspol.uw.edu.pl/polniem/), and others. Written corpora of historical Polish are also being actively developed. The largest monitor corpus of Polish is Monco PL (monco.frazeo.pl) with over 7.7 billion words and a collocation search (Pęzik 2020). The recently released ELEXIS Polish Web corpus is currently the largest corpus, with over 12 billion tokens.

All of the corpora mentioned above are freely searchable but due to copyright issues they cannot be freely downloaded and further used for language technology

[1] Based on Ogrodniczuk/Łaziński/Miłkowski/Pęzik (in print).

processing. Some small corpora, such as the Polish Corpus at Wrocław University of Technology, Open Subtitles (film subtitles in Polish), Wolne Lektury (Free Lecture) are freely distributable but not balanced and not up-to-date. The ELEXIS corpus is freely downloadable for research purposes because it contains only public web documents but it is not balanced either. A list of over 200 resources and tools for Polish can be found at: http://clip.ipipan.waw.pl/LRT.

The National Corpus of Polish is a basic resource for research in the humanities and the testbed for developing many language technology tools, including the first of their kind for Polish: morphological analyzers, disambiguating taggers or named entity recognizers.

Apart from the National Corpus of Polish, another project which has significantly changed the state of Polish language technology is CLARIN-PL, the Polish part of the pan-European Common Language Resources & Technology Infrastructure aimed at researchers in the humanities and social sciences. The co-operation of many research institutions led to the development of many language technology resources and tools such as:

- Słowosieć, the Polish WordNet, a relational lexico-semantic dictionary of Polish with almost 200 thousand lexemes and 600 thousand lexical relations (Dziob et al. 2019),
- Korpusomat, a corpus creation tool for non-technical users: https://korpusomat.pl/ (Kieraś/Kobyliński 2021),
- COMBO, a neural tagger, lemmatizer and dependency parser (Rybak/Wróblewska 2018),
- SpokesPL – a search engine for Polish conversational data: http://spokes.clarin-pl.eu/.

The development of language technology in Poland is based on four pillars:

- Research labs and groups mainly located at universities and the institutes of the Polish Academy of Sciences,
- Government-based institutions and ministries responsible for drafting strategic documents,
- Companies, both the big international players as well as mid-size companies and startups,
- Independent researchers, without any formal affiliation, often forming informal research groups gathered around meetups.

Linking Language Technology and Natural Language Processing to Artificial Intelligence has already happened in Poland with the advent of deep neural network powered solutions but its consequences are more far reaching than we can imagine. However, even when the technology seems mature enough, its absorption by larger public institutions and companies is proceeding much more slowly.

The availability of deep neural network powered frameworks has moved the focus from tools to resources. Therefore, an awareness of the value of data is still increasing. This includes opening up public data and eliminating legal barriers to the exploration of Polish data under copyright protection.

A crucial, and maybe the most important, factor in the development of Polish Language Technology is the support of the national research community with international cooperation. Polish Language Technology research has already benefited from numerous pan-European initiatives such as ELRC (European Language Resource Coordination, https://lr-coordination.eu/), ELG (European Language Grid, https://www.european-language-grid.eu/) and ELE (European Language Equality, https://european-language-equality.eu/), research infrastructures such as CLARIN and DARIAH, COST Actions and CEF projects. This trend must continue to strengthen the European research community.

References

Dziob, A./Piasecki, M./Rudnicka, E. (2019): plWordNet 4.1 – a linguistically motivated, corpus-based bilingual resource. In: *Proceedings of the 10th Global Wordnet Conference*. Wroclaw, 353–362. https://aclanthology.org/2019.gwc-1.45.

Kieraś, W./Kobyliński, Ł. (2021): Korpusomat – stan obecny i przyszłość projektu. In: *Język Polski* CI (2), 49–58.

Ogrodniczuk, M./Pęzik, P./Łaziński, M./Miłkowski, M. (in print): *European Language Equality. D.1.217 Report on the Polish Language*. European Language Resource Coordination.

Pęzik, P. (2020): Budowa i zastosowania korpusu monitorującego MoncoPL. In: *Forum Lingwistyczne* (7), 133–150. http://doi.org/10.31261/FL.2020.07.11.

Przepiórkowski, A./Bańko, M./Górski, R. L./Lewandowska-Tomaszczyk, B. (eds.) (2012): *Narodowy Korpus Języka Polskiego*. Warsaw.

Rybak, P./Wróblewska, A. (2018): Semi-supervised neural system for tagging, parsing and lematization. In: *Proceedings of the CoNLL 2018 – Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, 45–54. http://www.aclweb.org/anthology/K18-2004.

Elena Isabelle Tamba

The role of the Institutes of the Romanian Academy in the digitalization process of linguistic research

Abstract

In the last few years, measures have been taken in Romania to create the necessary electronic instruments and resources to support the Romanian language and culture on a transnational level in the general context of the digitalization of basic academic research. In today's digital, multicultural society, this had become an absolutely necessary step to take.

Electronic dictionaries and text corpora structured as databases facilitate knowing, preserving and maintaining cultural identity on a linguistic level and allow the inclusion of a national language in the field of interest of digitalized research into natural languages on a global level.

1. Introduction

One of the objectives of European policies is the preservation and valorization of national linguistic identities, as long as there is a general tendency towards using languages which are privileged by the existence of (electronic) means of promoting them.

Linguistics and lexicography around the world have undergone an extensive process of change, including the modernization of means of writing, consulting, etc., through approaches that involve interconnections between different fields of research.

Romanian linguistics and lexicography have also been marked by this change. In the last few years, measures have been taken in Romania, to create the necessary electronic instruments and resources to support the Romanian language and culture on a transnational level in the general context of the digitalization of basic academic research.

A special stage in the evolution of Romanian linguistics and lexicography at present is the digitalization of research, which involves the digitalization of existing resources on the one hand and digitalization – the creation of dictionaries, new resources, and instruments in an electronic format – on the other. In parallel, linguistic and lexicographic resources continue to be created in classical, printed format.

Basically, digitalization involves converting existing resources in printed format into an electronic format. For example, linguistic and lexicographic corpora, can be created by digitizing printed dictionaries; linguistic corpora can be annotated morphologically, syntactically, and semantically and lexicographic corpora can include digitalized dictionaries.

Digitalization, in turn, involves the development of lexicographic or linguistic resources such as dictionaries directly in electronic format by creating/using dictionary writing programs and/or sample extraction programs, etc.

Most efforts in the digitalization of lexicographic research have been made under the auspices of the Institutes of the Romanian Academy; more recently, research centers at some universities in the country have become involved.

Today various digital linguistic/lexicographic projects are being carried out in Romania including:

- academic initiatives (most lexicographic digitalization projects are taking place at the Institutes for the Romanian Language and the IT Institutes of the Romanian Academy while a few are being developed at the research centers of some universities in Romania[1] or in some libraries).
- private initiatives (for example, https://dexonline.ro/ – a lexicographic platform initiated by volunteers, projects at some publishing houses, etc.).

In this paper we will highlight the projects of the Institutes of the Romanian Academy.

2. Digitalized linguistics and lexicography in Romanian

2.1 Institutes for Language at the Romanian Academy

In the Romanian Academy there are three institutes where research into the Romanian language is done in the fields of lexicography, lexicology, grammar, history of the language, dialectology, sociolinguistics, and onomastics, etc.:

- Institutul de Filologie Română “A. Philippide”, Iaşi/“A. Philippide” Institute of Romanian Philology – https://www.philippide.ro/,

[1] Here we would like to mention some lexicographic digital projects developed in two universities in Romania: *The Lexicon from Buda (1825). Amended and electronically processed edition for online consultation* (http://www.bcucluj.ro/lexiconuldelabuda/site/login.php), a project coordinated by the “Babeş-Bolyai” University of Cluj-Napoca and *Primele dicţionare bilingve româneşti* (*secolul al XVII-lea*). *Corpus digital prelucrat şi aliniat* (eRomLex) [*The first Romanian bilingual dictionaries* (*17th century*). *Digitally annotated and aligned corpus*. eRomLex] – the main objective of this project, developed at the “Alexandru Ioan Cuza” University of Iasi, is the elaboration of a comparative digital edition of the Slavonic Romanian dictionaries from the 17th century (all of them are manuscripts) – http://www.scriptadacoromanica.ro/bin/view/eRomLex/.

- Institutul de Lingvistică “Iorgu Iordan – Alexandru Rosetti”, Bucureşti/“Iorgu Iordan – Alexandru Rosetti” Institute of Linguistics – https://www.lingv.ro/,
- Institutul de Lingvistică şi Istorie Literară “Sextil Puşcariu”, Cluj/“Sextil Puşcariu” Institute of Linguistics and Literary History – http://www.inst-puscariu.ro/.

The main projects involving basic research concern the following reference works:

- Dictionary of the Romanian Language,
- Grammar of the Romanian Language,
- History of the Romanian Language,
- Linguistic Atlases covering different areas for the Romanian Language, etc.

Researchers from the above-mentioned institutes are also involved in some international projects, like: DERom (*Dictionnaire Étymologique Roman*, http://www.atilf.fr/DERom/), ENeL (*European Network of e-Lexicography*, https://www.elexicography.eu), ALE (*Atlas linguarom Europae* – https://lingv.ro/atlas-linguarum-europae/), etc.

Research into the Romanian language is also carried out at the IT Institutes of the Romanian Academy, namely in the fields of natural language processing or computational linguistics:

- Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu”, Bucureşti/“Mihai Drăgănescu” Research Institute for Artificial Intelligence – http://www.racai.ro,
- Institutul de Informatică Teoretică, Iaşi/Institute of Theoretical Informatics – http://iit.academiaromana-is.ro/.

2.2 Thesaurus Dictionary of the Romanian Language in the digital age

Great European cultures have had thesaurus dictionaries and text corpora in electronic format for many years now. The main Romanian lexicographic project is the *Thesaurus Dictionary of the Romanian Language* (DA/DLR), which is edited by the Romanian Academy and was started 115 years ago. That is why creating an electronic format which is accessible to scientists and everybody who is interested in learning or studying Romanian in our country or abroad became an absolutely necessary step to take in today’s digital, multicultural society.

Fig. 1: *Thesaurus Dictionary of the Romanian Language* (DA/DLR)

For a better understanding of the dimensions of the *Thesaurus Dictionary of the Romanian Language*, we present some statistics and compare them to other large European dictionaries:

- The first edition of the *Thesaurus Dictionary of the Romanian Language* was published in two series: DA (1907-1944) and DLR (1965-2010). It includes 14 tomes with 37 volumes, 20,000 lexicon type pages (between 7,000 and 11,000 characters per page), over 175,000 words (with variants) and over 1,300,000 quotes; the electronic form is being elaborated (first attempt 2007-2010; work in progress). The second edition is also work in progress.
- *Diccionario de la lengua espanola de la Real Academia Espagnola* (DRAE, https://dle.rae.es/diccionario): first printed edition – 1780; 23rd edition – 2014; 93,111 lemmas; first electronic format – 1992.
- *Dictionnaire de l'Académie Française* (https://dictionnaire-academie.fr/): first printed edition – 1694; 9 editions; available online, 55,000 words.
- *Deutsches Wörterbuch der Grimm* (DWB, http://germazope.uni-trier.de/Projects/DWB): 1838-1961; 32 volumes; 350,000 words and variants; first electronic format: 1997-2004.
- *Oxford English Dictionary* (OED, http://www.oed.com/): first edition – 1928, 20 volumes; second edition – 1989; 301,100 words, 2,412,400 quotes; first electronic format – 1988.
- *Tresor de la Langue Française* (TLF), XIXth-XXth centuries (http://atilf.atilf.fr/): first printed edition – 1971-1994; 16 volumes; 100,000 words, 270,000 definitions, 430,000 quotes; electronic format: 1990-2004.
- *Tesoro della lingua italiana delle origini* (TLIO, http://tlio.ovi.cnr.it/TLIO/index2.html): 44,000 words (37,864 published online) out of an intended 57,000 words.

Based on the data above, the *Thesaurus Dictionary of the Romanian Language* can be compared, both in terms of its conception and realization, with similar dictionaries of European languages, and its digitalization is, thus, a normal step in the evolution of Romanian lexicography.

We are preparing the digital form of the *Dictionary* in three projects:

- digitalization of the printed form in the **eDTLR** project (scanning, OCR correction, correction, parsing and uploading to a platform which allows complex searches in the body of each lexicographic entry);
- digitalization of the printed form in the **CLRE** project (scanning and processing in the CLRE platform, which allows, for the time being, consultations at headword level and displaying an image of the page from the dictionary);
- digitalization of the **second edition** of the **DLR** (editing done entirely and directly in a dictionary-writing program).

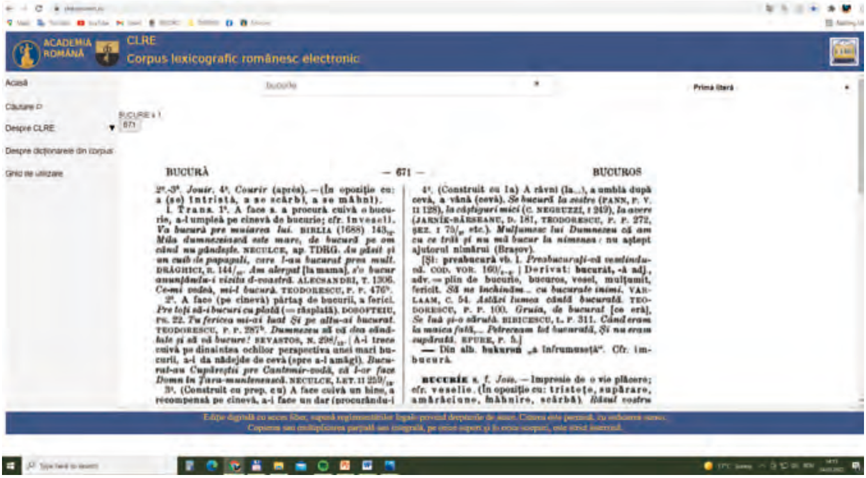
Digitalization of the Dictionary started in 2007 (until 2010), in a complex project eDTLR *Dicționarul tezaur al limbii române în format electronic* (*Romanian Thesaurus Dictionary in electronic format*) which had as its main objective the acquisition of the complete form of the *Thesaurus Dictionary of the Romanian Language* into electronic format as a result of retro-digitalization, but the research is continuing. The results of the eDTLR project will make the electronic format of the *Thesaurus Dictionary of the Romanian Language* accessible for everybody who knows or is interested in Romanian. The digital form of this *Dictionary of the Romanian Language* in CLRE will be presented in the next section.

The second edition of the *Thesaurus Dictionary of the Romanian Language* is called DLRi (*Dicționarul Limbii Române informatizat – Digital Dictionary of the Romanian Language*). It was started in 2010 by electronically acquiring the textual resources of the Bibliography. (The Romanian language does not have yet a complete electronic corpus – it is still work in progress.) We are now working with an electronic editing interface, adapted by Oxygen. The DLRi is being developed completely in electronic form. A printed format will also be published in parallel. The first part of the letter A was presented to the public in digital format in May 2021 – https://dlri.ro/.

2.3 CLRE: The Essential Romanian Lexicographic Corpus

Creating an Electronic Romanian Lexicographic Corpus has been a constant concern of Romanian lexicographers in the last fifteen years, a fact justified by the broader context of the digitalization of Romanian research.

CLRE. *Corpus lexicografic românesc electronic* (***CLRE***. *Electronic Romanian Lexicographic Corpus*) is a project carried out by the Romanian Academy which involves an electronic collection of dictionaries of Romanian aligned at the entry level. It includes the most important lexicographic works from the very first one written in Romanian in the 17th century to the latest ones. The corpus includes, as its main lexicographic work, the *Thesaurus Dictionary of the Romanian language* (DA/DLR).



BUCURĂ — 671 — BUCUROS

2°-3°. *Jouir*. 4°. *Courir* (après). — (În opoziție cu: a (se) întristà, a se scârbì, a se mâhnì).

I. Trans. 1°. A face s. a procurà cuivà o bucurie, a-l umpleà pe cinevà de bucurie; cfr. înveselì. *Va bucurà pre muiarea lui*. BIBLIA (1688) 143. *Mila dumnezeiască este mare, de bucură pe om când nu gândeşte*. NECULCE, ap. TDRG. *Am găsit şi un cuib de papagali, care l-au bucurat prea mult*. DRĂGHICI, R. 144/. *Am alergat* [la mama], *s'o bucur anunţându-i vizita d-voastră*. ALECSANDRI, T. 1306. *Ce-mi vedeà, mi-l bucurà*. TEODORESCU, P. P. 476.

2°. A face (pe cinevà) părtaş de bucurii, a ferici. *Pre toţi să-i bucuri cu plată* (= răsplată). DOSOFTEIU, PS. 22. *Tu ferice mi-ai luat Şi pe altu-ai bucurat*. TEODORESCU, P. P. 287. *Dumnezeu să vă dea sănătate şi să vă bucure!* SEVASTOS, N. 298/. | A-i trece cuivà pe dinaintea ochilor perspectiva unei mari bucurii, a-i da nădejde de cevà (spre a-l amăgi). *Bucuratu-au Cuparveştii pre Cantemir-vodă, că l-or face Domn în Ţara-munteneacă*. NECULCE, LET. II 259/.

3°. (Construit cu prep. cu) A face cuivà un bine, a recompensà pe cinevà, a-i face un dar (procurându-i

4°. (Construit cu la) A râvnì (la...), a umblà după cevà, a vânà (cevà). *Se bucură la zestre* (PANN, P. V. II 128), *la câştiguri mici* (C. NEGRUZZI, I 249), *la avere* (JARNÍK-BÂRSEANU, D. 181, TEODORESCU, P. P. 272, ŞEZ. I 75/ etc.). *Mulţumesc lui Dumnezeu că am cu ce trăi şi nu mă bucur la nimenea*: nu aştept ajutorul nimărui (Braşov).

[Şi: preabucurà vb. I. *Preabucuraţi-vă vestindu-vă*. COD. VOR. 160/. | Derivat: bucurát, -ă adj., adv. = plin de bucurie, bucuros, vesel, mulţumit, fericit. *Să ne închinăm... cu bucurate inimi*. VARLAAM, C. 54. *Astări lumea cântă bucurată*. TEODORESCU, P. P. 100. *Gruia, de bucurat [ce eră], Se luă şi-o sărută*. BIBICESCU, L. P. 311. *Când eram la maica fată,... Petreceam tot bucurată, Şi nu eram supărată*. EPURE, P. 5.]

— Din alb. bukuronj „a înfrumuseţà". Cfr. îmbucurà.

BUCURÍE s. f. *Joie*. — Impresie de o vie plăcere; cfr. veselie. (În opoziţie cu: tristeţe, supărare, amărăciune, mâhnire, scârbă). *Râsul vostru*

Ediție digitală cu acces liber, supusă reglementărilor legale privind drepturile de autor. Citarea este permisă, cu indicarea sursei.
Copierea sau multiplicarea parțială sau integrală, pe orice suport și în orice scopuri, este strict interzisă.

Fig. 2: *Corpus lexicografic românesc electronic* (CLRE)/*Electronic Romanian Lexicographic Corpus* (CLRE)

The main objectives of the CLRE project are:

- to create the largest digital diachronic corpus of dictionaries of Romanian consisting of lexicographic works from the digitized DLR Bibliography (transposed from its classical format, on paper, into digital format) and from digitalized dictionaries (created in an editable electronic format);
- to promote lexicographic works produced under the auspices of the Romanian Academy;
- to provide information from CLRE with free access for the general public.

The first work chosen by the lexicographers from Iasi for publication in CLRE is the *Dictionary of the Romanian Language* produced by the Romanian Academy. The first volume was digitized and published online in September 2021 as Volume I. Part I: A-B and contains 8,517 entries (https://clre.solirom.ro/). This choice was justified by the fact that this is the first volume of the *Dictionary of the Romanian Language* published under the auspices of the Academy and by its parallelism with the publication of the first part of the second edition of DLRi, letter A, written by fellow lexicographers from the Institute of Linguistics "Iorgu Iordan – Al. Rosetti", Romanian Academy, Bucharest (https://dlri.ro/).

CLRE can be compared to two other European lexicographic corpora which are similar in their technical approach:

- *Diccionarios de la lengua española* – a database containing dictionaries edited and published by the Real Academia Espagnola (https://www.rae.es/obras-academicas/diccionarios).

- *Das Wörterbuchnetz* – a collection of 37 electronic dictionaries created at the University of Trier in Germany (https://www.woerterbuchnetz.de/).

The development of CLRE, mirroring other directions for the development of electronic resources, represents a starting point for future research, which may be part of a medium- and long-term research strategy, such as:

- aligning the *Romanian Thesaurus Dictionary* in electronic format (eDTLR) with CLRE DA/DLR and other dictionaries from the corpus;
- using CLRE to elaborate the DLRi (the second edition) and for other lexicographic projects;
- developing large-scale applications on the semantic disambiguation of words;
- selecting entry types to produce new, specialized dictionaries (thematic, etymological, etc.);
- highlighting dictionaries from the database by publishing them online or republishing a dictionary in a mixed format (classical and online);
- turning CLRE into an open corpus (in the sense of the possibility of adding new lexicographic works) for all researchers from the Romanian Academy;
- associating it with other linguistic or multimedia resources, which would bring Romanian lexicography to a level comparable with European lexicography (for example, with the *Dictionnaire Étymologique Roman* (DÉRom) (http://www.atilf.fr/DERom/) or ENeL: European Network of e-Lexicography (http://www.elexicography.eu).

2.4 TDRG

Another lexicographic project published online by the Romanian Academy is the electronic version of the **TDRG** – H. Tiktin, *Rumänisch-Deutsches Wörterbuch* (first edition 1896-1926). The third edition of this dictionary (published in 2003-2005) was digitalized, following the model of eDTLR, in a project involving the Albert-Ludwigs-Universitat in Freiburg, Germany, and the Romanian Academy, and it has now been published online (https://tdrg.solirom.ro/).

2.5 SOLIROM

All of the results of the digitalization process of linguistic/lexicographic research in the Institutes for Language of the Romanian Academy are planned to be published together online on an academic platform called **SOLIROM** (https://solirom.ro/). It will include all electronic resources (DLRi, CLRE, TDRG, eDTLR, etc.) either directly or via a link to the homepage of the project.

Until last year, every academic project mentioned above was published online on a separate web page, but now the results of these projects (digitalized or digital dictionaries) have been published (or will be published) online on this single platform of the Romanian Academy. SOLIROM promotes a unitary way of working, at the level of specialized institutes, regarding the creation of digital linguistic resources and tools dedicated to the Romanian language and literature. Important projects of the Romanian Academy, such as the new digital edition of the *Dictionary of the Romanian Language* involves permanent collaboration between teams of researchers from several institutions in the country, the use of the same documentation sources and writing tools, so the approach offered by the SOLIROM platform is welcome. This allows, among other things, the alignment of developed language resources, the simplification and streamlining of the publishing process using website templates, as well as the management of published digital resources with minimal resources, which is an important element in the management of research activity.

The platform consists of two sections, a public one which provides digital language resources for public access and a private one with the digital tools needed to manage the platform's digital language resources for the researchers developing the platform.

Now the Romanian Academy is developing a new site with a special area dedicated to Romanian language resources.

2.6 CoRoLa

Another very important project concerning digital resources for Romanian is ***Corpus computațional de referință pentru limba română contemporană*** [Reference computational corpus for contemporary Romanian language] **– CoRoLa** (http://corola.racai.ro/).[2]

The purpose of CoRoLa is to be an online resource for the study and learning of Romanian and so it is a very important resource for lexicographic research as well.

Starting in 2014, this corpus was developed as a priority program of the Romanian Academy. It contains various texts, dating from 1989 to the present day, the purpose of its creation being to provide an objective image of current written and spoken Romanian. The corpus is publicly accessible via two interfaces, one for searching for text data and one for searching for audio data. The main fields of use of the CoRoLa corpus are: linguistic studies; language modeling for the automatic processing of Romanian; developing translation models; language learning; intelligent and multi-criteria indexing and retrieval of textual and oral

[2] Another online resource related to the Romanian Academy is **DIGIBUC** (http://www.digibuc.ro/), the most important Romanian digital library, a project run by the Bucharest Metropolitan Library and the Library of the Romanian Academy. It is the official partner of the European Digital Library EUROPEANA (http://www.europeana.eu/portal/).

information; semantic classification of large volumes of data (text and audio); extracting knowledge from data (text and audio); automatic document summaries; question-answer systems; automatic speech recognition and synthesis; and so on.

3. Conclusions

The aim of this paper is to highlight, in general, the current status of linguistic and lexicographic research in the Institutes of The Romanian Academy in the digital age.

Trends in Romanian linguistics and lexicography include:

- Writing online dictionaries based on continuously increasing text corpora and on various tools (programs for extracting the quotations, for example);
- Developing a Romanian Language Text Corpora (for Contemporary Romanian we have the CoRoLa corpus; a diachronic corpus – work in progress), and linking it to the Thesaurus Dictionary;
- Developing lexicographic corpora (CLRE – work in progress);
- Using dictionary writing systems (DLRi – work in progress);
- Further editing of the printed edition of the *Thesaurus Dictionary of the Romanian Language*;
- Aligning various lexicographic works and creating collaborative programs between academics with lexicographically-oriented publishers etc., as an important subsequent goal;
- Matching electronic lexicographic resources for Romanian – DLRi – CLRE – eDTLR etc. – and all of them with other linguistic resources (possibly multimedia) from Romania and abroad.

Electronic dictionaries and text corpora structured as databases facilitate knowing, preserving, and maintaining cultural identity on a linguistic level and allow the inclusion of a national language in the field of interest of digitalized research into natural languages on at a global level.

References

Ernst, G. (2013): “Romanian”. In: Heid, U./Gouws, R.H./Schweickard, W./Wiegand, H.E. (eds.): *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*, Berlin/Boston, 687-701.

Hartmann, R.R.K./James, G. (1998): *Dictionary of lexicography*, London.

Kirchmeier, S. (2020): Trends in European language policies with a view to language technology. In: *Bendrinė* Kalba 93. http://journals.lki.lt/bendrinekalba.

Tamba, E. (2014): La lexicografía Rumana. Historia y Actualidad. In: Córdoba Rodríguez, F./González Seoane, E./Sánchez Palomino, M.D. (eds.): *Lexicografía de las lenguas románicas. Perspectiva histórica. Vol. I*. Berlin/Munich/Boston, 265-282.

Tamba Dănilă, E./Clim, M.-R./Catană-Spenchiu, A./Patraşcu, M. (2012): The evolution of the Romanian digitalized lexicography. The Essential Romanian Lexicographic Corpus. In: Vatvedt Fjeld, R./Torjusen, J.M. (eds.): *Proceedings of the 15th EURALEX International Congress, 7-11 August 2012*. Oslo, 1014-1017. http://www.euralex.org/proceedings-toc/euralex_2012/.

Tamba, E.I. (2017): CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din Bibliografia DLR aliniate la nivel de intrare şi la nivel de sens. In: Haja, G. (ed.): *Lexicografia academică românească. Studii. Proiecte*. Iaşi, 221-234.

Dictionaries

DA = *Dicționarul limbii române*, tom I-II, Tipografia ziarului "Universul". Bucharest, 1907-1944.

DAF = *Dictionnaire de l'Académie Française*. https://dictionnaire-academie.fr/.

DLR = *Dicționarul limbii române (DLR)*, Serie nouă, tom. VI-XIV, Bucharest, 1965-2010.

DRAE = *Diccionario de la lengua espanola de la Real Academia Española*. http://buscon.rae.es/draeI/.

DWB = *Deutsches Wörterbuch "der Grimm"*. http://germazope.uni-trier.de/Projects/DWB.

OED = *Oxford English Dictionary*. http://www.oed.com/.

TLFi = *Le Trésor de la Langue Française Informatisé*. http://atilf.atilf.fr/.

TLIO = *Tesoro della lingua italiana delle origini*. http://tlio.ovi.cnr.it/TLIO/index2.html.

Kozma Ahačič/Nataša Gliha Komac/Janoš Ježovnik

Developing a comprehensive service for Slovenian language users: the Fran and Franček web portals and language advisory service

Abstract (English)

Since 2014, the ZRC SAZU Fran Ramovš Institute of the Slovenian Language has successfully established a comprehensive service for Slovenian language users. Children and teens can obtain information on nearly 100,000 words via the innovative web portal Franček.si which has been adapted to their needs while general and professional users can simultaneously browse up to forty Slovenian dictionaries on the Fran web portal. In addition, we maintain regular contact with Slovenian language users via two advisory pages: one targeting the general public and one aimed at developing terminology in a multilingual society.

Abstract (Slovenian)

Na Inštitutu za slovenski jezik Frana Ramovša ZRC SAZU smo v času od leta 2014 do danes uspeli vzpostaviti popoln servis za uporabnike slovenskega jezika. Mladi lahko dobijo podatke o skoraj 100.000 besedah na inovativnem, njim prilagojem portalu Franček.si, splošni in profesionalni uporabniki lahko hkrati brskajo po kar 40 slovarjih slovenskega jezika na portalu Fran, hkrati pa smo v stalnem stiku z uporabniki slovenskega jezika prek dveh svetovalnic: ena je namenjena splošni javnosti, druga pa razvoju terminologije v večjezični družbi.

1. The Fran web portal and language advisory services

In 2014, the Fran Ramovš Institute of the Slovenian Language at ZRC SAZU found itself at an important crossroads. The institute had financial problems and we were unable to meet the new demands of language users at that time. Our online dictionaries could only be accessed one at a time and they did not have a uniform format.[1] Therefore, we decided to completely rework our approach for users of our language manuals.

We combined all major lexicographic sources in one uniform portal system, taking advantage of the fact that practically all basic lexicographic works and sources related to Slovenian had been created at the institute. This allowed us to

[1] Cf. the old website: http://bos.zrc-sazu.si.

focus primarily on content-related issues from the very beginning rather than on time-consuming legal procedures to acquire copyright for publishing the works.

The year 2014 saw the launch of the comprehensive Slovenian dictionary portal Fran.si (see Perdih 2018, 2020), which brings together all dictionary entries, sources, and materials created by the institute's researchers. The portal, which has recorded an average of over 300,000 searches a day over the past year (bearing in mind that Slovenia's population is just over two million),[2] provides access to twelve general dictionaries,[3] two etymological dictionaries,[4] five historical dictionaries,[5] fourteen terminological dictionaries,[6] five dialect dictionaries,[7] a dialect atlas,[8] and two language advisory services, one intended for general users[9] and one intended for specialists in various fields.[10]

The search engine allows users to conduct both simple and complex searches in all dictionaries at once (see Fig. 1); their indexing system allows extremely fast access to as many as 689,941 dictionary entries in various reference works. The order of the displayed search results adapts to users' interests on an ongoing basis (see Fig. 2), making the portal easy to use even for those not familiar with different types of dictionaries, their structures, and their purpose.

[2] https://www.fran.si/o-portalu?page=Statistics.

[3] Dictionary of the Slovenian Standard Language, 2nd Edition, Dictionary of the Slovenian Standard Language, 3rd Edition (eSSKJ; 2016–), Synonym Dictionary of Slovenian Language, Slovenian Normative Guide, ePravopis – Slovenian Normative Guide (2016–), Growing Dictionary of the Slovenian Language (2014–), Dictionary of Slovenian Phrasemes, Dictionary of Slovenian Valency, Dictionary of Proverbs and Similar Paremiological Expressions, Dictionary of New Slovenian Words etc.

[4] Slovenian Etymological Dictionary, NESSJ – New Etymological Dictionary of Slovenian Language (2017–).

[5] Words of the 16th-Century Slovenian Literary Language, Dictionary of the Slovenian Language in the Works of John Baptist of Sveti Križ, Dictionary of Kastelec and Vorenc (1680-1710), Dictionary of the Language of Marko Pohlin, Dictionary of Old Standard Prekmurje Slovenian, Dictionary of Maks Pleteršnik (1894-1895).

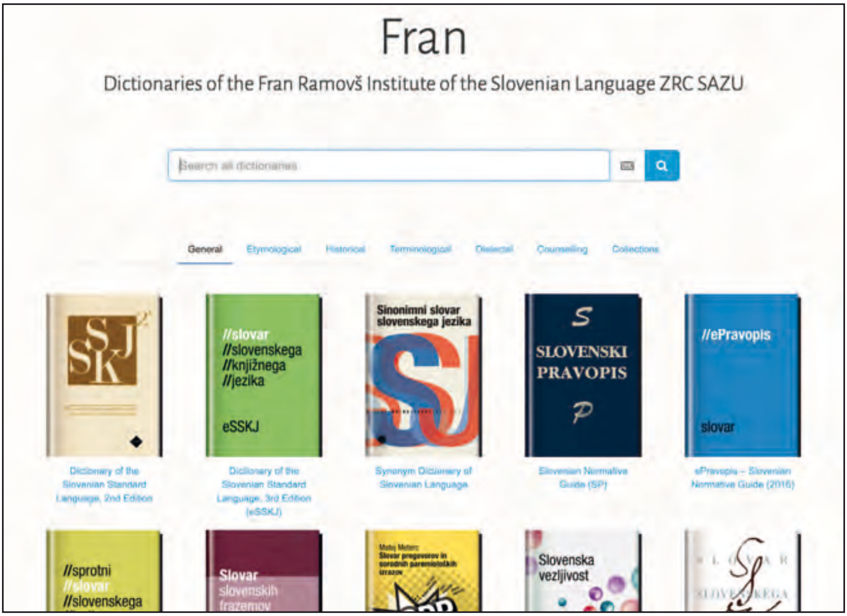
[6] Concrete Structures, Pharmacy, Law, Automatic Control, Systems and Robotics, Urban Planning, Applied Art, Percussion, Botany, Skiing, Theatre, Beekeeping, Geology, Gemology, Geography, Mountaineering.

[7] Dictionary of the Črni Vrh Dialect, Dictionary of the Local Dialects of the Dreta Valley, Dictionary of the Bovec Local Dialect, Dictionary of the Kostel Dialect, Dictionary of the Clothing Terminology of the Gail Valley, Local Dialect of the Canal Valley.

[8] https://www.fran.si/204/sla-slovenski-lingvisticni-atlas.

[9] https://svetovalnica.zrc-sazu.si.

[10] https://isjfr.zrc-sazu.si/sl/terminologisce/svetovanje.



Fran

Dictionaries of the Fran Ramovš Institute of the Slovenian Language ZRC SAZU

Search all dictionaries

General · Etymological · Historical · Terminological · Dialectal · Counselling · Collections

Dictionary of the Slovenian Standard Language, 2nd Edition

Dictionary of the Slovenian Standard Language, 3rd Edition (eSSKJ)

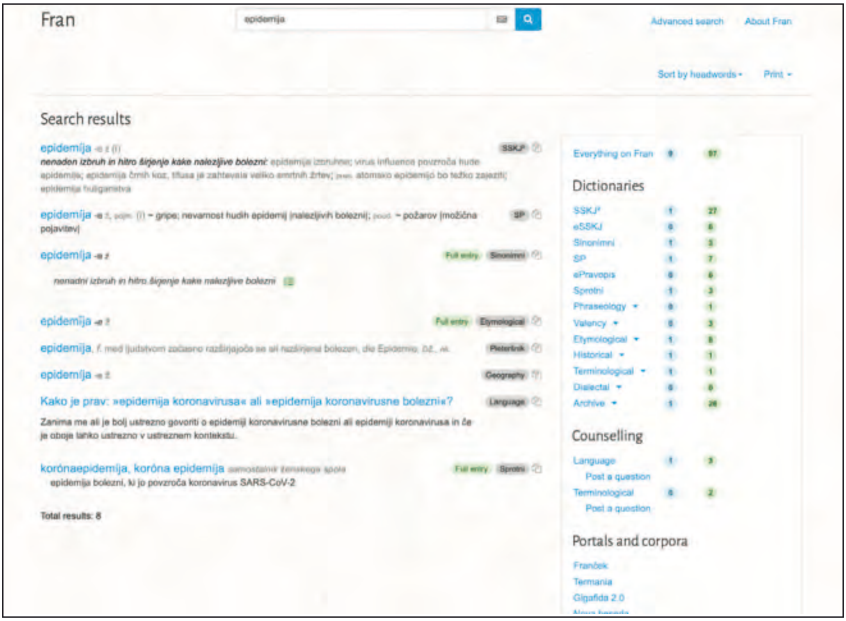
Synonym Dictionary of Slovenian Language

Slovenian Normative Guide (SP)

ePravopis – Slovenian Normative Guide (2016)

SSKJ²
//slovar //slovenskega //knjižnega //jezika
eSSKJ
Sinonimni slovar slovenskega jezika
S
SLOVENSKI PRAVOPIS
P
//ePravopis
slovar
//sprotni //slovar //slovenskega
Slovar slovenskih frazemov
Matej Meterc
Slovar pregovorov in sorodnih paremioloških izrazov
Slovenska vezljivost
SLOVAR SLOVENSKEGA

Fig. 1: The Fran web portal: entry page



Fran

epidemija

Advanced search · About Fran

Sort by headwords ▾ · Print ▾

Search results

epidemija -e ž (I) — SSKJ²
nenadon izbruh in hitro širjenje kake nalezljive bolezni: epidemija izbruhne; virus influence povzroča hude epidemije; epidemija črnih koz, tifusa je zahtevala veliko smrtnih žrtev; *pren.* atomsko epidemijo bo težko zajeziti; epidemija huliganstva

epidemija -e ž, pojm. (I) ~ gripe; nevarnost hudih epidemij |nalezljivih bolezni|; *pod.* ~ požarov |množična pojavitev| — SP

epidemija -e ž — Full entry · Sinonimni
nenadni izbruh in hitro širjenje kake nalezljive bolezni

epidemīja -e ž — Full entry · Etymological

epidemija, *f.* med ljudstvom začasno razširjajoča se ali razširjena bolezen, die Epidemie, *Dž.*, *nk.* — Pleteršnik

epidemija -e ž — Geography

Kako je prav: »epidemija koronavirusa« ali »epidemija koronavirusne bolezni«? — Language
Zanima me ali je bolj ustrezno govoriti o epidemiji koronavirusne bolezni ali epidemiji koronavirusa in če je oboje lahko ustrezno v ustreznem kontekstu.

korónaepidemija, koróna epidemija samostalnik ženskega spola — Full entry · Sprotni
epidemija bolezni, ki jo povzroča koronavirus SARS-CoV-2

Total results: 8

Everything on Fran	8	97
Dictionaries		
SSKJ²	1	27
eSSKJ	0	0
Sinonimni	1	3
SP	1	7
ePravopis	0	6
Sprotni	1	3
Phraseology ▾	0	1
Valency ▾	0	3
Etymological ▾	1	8
Historical ▾	1	1
Terminological ▾	1	1
Dialectal ▾	0	0
Archive ▾	1	28
Counselling		
Language	1	3
Post a question		
Terminological	0	2
Post a question		

Portals and corpora

Franček
Termania
Gigafida 2.0
Nova beseda

Fig. 2: The Fran web portal: search results

The Fran portal also reacts to its users' queries. Based on search analyses and questions addressed to the general language and terminological advisory service webpages, a dictionary of neologisms[11] is being compiled to respond to user needs as they are expressed. For example, during the COVID-19 pandemic we were already able to post explanations of the most frequently used neologisms within a month of the outbreak and we then updated these data on an ongoing basis.

Thanks to Fran's flexibility, the dictionaries produced at the institute can be posted on the portal as works in progress (that is, even before they are completed). The missing dictionary entries are simply covered by those in existing dictionaries.

Users are always told which dictionary they are currently looking at and can quickly access information on that dictionary's characteristics and specific features.[12]

They can send us their questions by e-mail or via a dedicated interface on the language advisory service webpages, and they can also send their suggestions directly via the online form on the Fran portal. In addition to corpus data, the so-called growing dictionaries (that is, dictionaries which are being developed and published in real time) can also rely on data on the most frequently searched words (see Fig. 4), the material on the two advisory pages, and users' direct suggestions and preferences. In this way we can also respond to the language needs of specific groups, such as the deaf, blind, and other vulnerable groups.

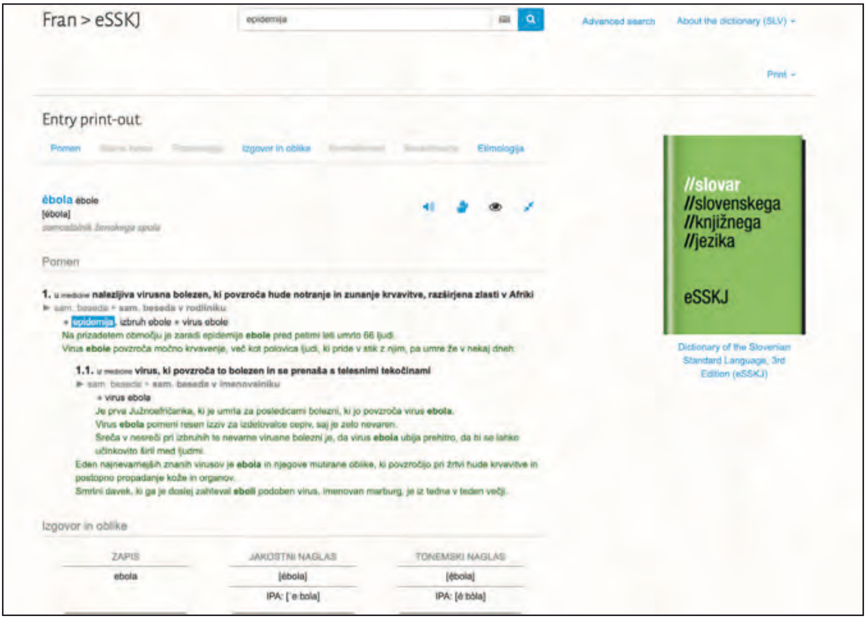
Another advantage of Fran is that it displays hits from both advisory pages alongside those from the dictionaries. This means that, in addition to a dictionary entry describing a specific word, users also receive information on special features of its contemporary use.

The problem of the specific role of terminological dictionaries is being solved by incorporating them in the Fran portal. In addition, specialized users can access them on the separate website *Terminologišče*, which provides various kinds of information on terminology, such as a selection of terminological articles and books with active links, various terminological dictionaries, and a terminological advisory service.[13]

[11] https://www.fran.si/132/sprotni-sprotni-slovar-slovenskega-jezika.

[12] https://www.fran.si/iskanje?View=1&Query=epidemija.

[13] https://isjfr.zrc-sazu.si/sl/terminologisce.



Fran > eSSKJ

epidemija

Advanced search About the dictionary (SLV)

Print

Entry print-out

Pomen Izgovor in oblike Etimologija

ébola ébole
[ébola]
samostalnik ženskega spola

Pomen

1. u medicini nalezljiva virusna bolezen, ki povzroča hude notranje in zunanje krvavitve, razširjena zlasti v Afriki
► sam. beseda + sam. beseda v rodilniku
• epidemija, izbruh ebole • virus ebole
Na prizadetem območju je zaradi epidemije ebole pred petimi leti umrlo 66 ljudi.
Virus ebole povzroča močno krvavenje, več kot polovica ljudi, ki pride v stik z njim, pa umre že v nekaj dneh.

1.1. u medicini virus, ki povzroča to bolezen in se prenaša s telesnimi tekočinami
► sam. beseda + sam. beseda v imenovalniku
• virus ebola
Je prva Južnoafričanka, ki je umrla za posledicami bolezni, ki jo povzroča virus ebola.
Virus ebola pomeni resen izziv za izdelovalce cepiv, saj je zelo nevaren.
Sreča v nesreči pri izbruhih te nevarne virusne bolezni je, da virus ebola ubija prehitro, da bi se lahko učinkovito širil med ljudmi.
Eden najnevarnejših znanih virusov je ebola in njegove mutirane oblike, ki povzročijo pri žrtvi hude krvavitve in postopno propadanje kože in organov.
Smrtni davek, ki ga je doslej zahteval eboli podoben virus, imenovan marburg, je iz tedna v teden večji.

Izgovor in oblike

ZAPIS	JAKOSTNI NAGLAS	TONEMSKI NAGLAS
ebola	[ébola]	[ẹ́bola]
	IPA: [ˈeːbola]	IPA: [éːbòla]

Dictionary of the Slovenian Standard Language, 3rd Edition (eSSKJ)

Fig. 3: The Fran web portal: a growing dictionary entry

2. The Franček web portal

Our latest achievement is the school web portal Franček (Francek.si) (see Ahačič et al. 2021; Perdih et al. 2021; Petric 2020). The challenge we encountered in designing it was how to introduce primary- and secondary-school students to using dictionaries on a daily basis. Franček is a school portal that gradually develops dictionary skills in young people and complements lexicographic information with grammatical information via links to an online grammar. Franček consists of a list of headwords, which includes nearly 100,000 entries, and is linked to various linguistic databases Perdih (2021).

The entries of individual words in the portal (see Fig. 5) are adapted for three age groups and cover semantics (“What does this word mean?”), synonymy (“Find words with similar meaning”), phraseology (“Which multi-word unit does this word appear in?”), pronunciation (“How do I pronounce this word?”), inflections (“How is this word inflected?”), dialect use (“How is this word used in dialects?”; users can even record how they pronounce a specific word in their dialect), etymology (“What is the origin of this word?”), and history (“Since when has this word been used?”). The entries differ from the traditional lexicographic presentation in that they approximate how a word is explained by a teacher in class or by parents at home. Every entry on Franček is linked to a dictionary entry in its original format on Fran.

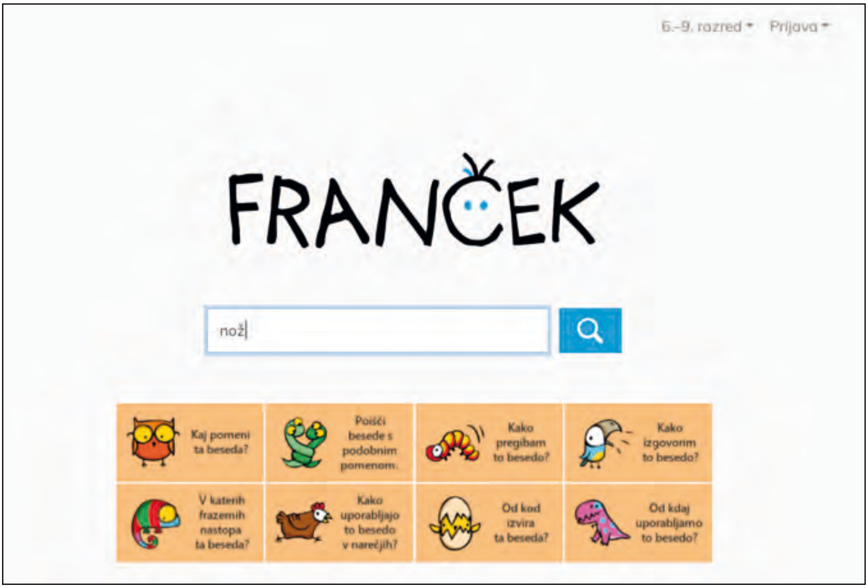
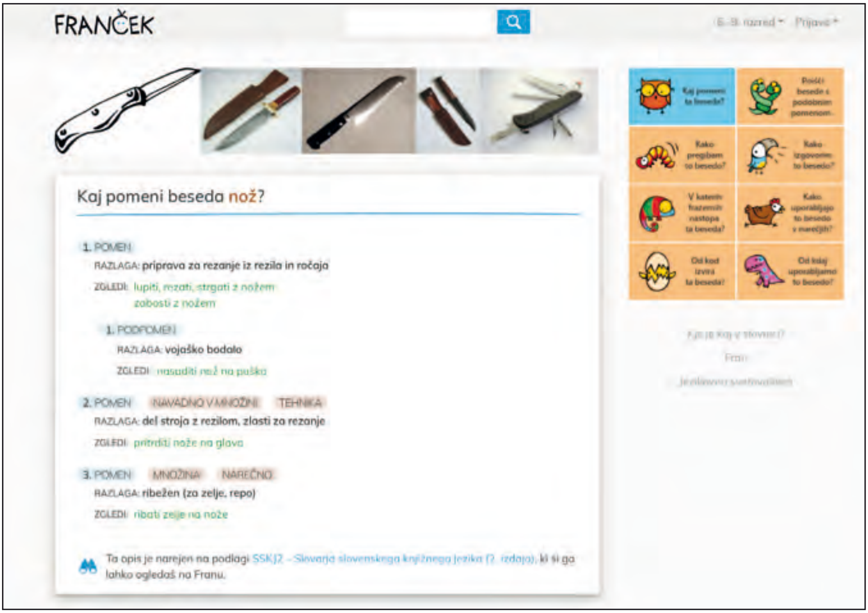

6.–9. razred
Prijava
FRANČEK
nož
Kaj pomeni ta beseda?
Poišči besede s podobnim pomenom.
Kako pregibam to besedo?
Kako izgovorim to besedo?
V katerih frazemih nastopa ta beseda?
Kako uporabljajo to besedo v narečjih?
Od kod izvira ta beseda?
Od kdaj uporabljamo to besedo?

Fig. 4: The Franček web portal: entry page


FRANČEK
6.–9. razred
Prijava
Kaj pomeni ta beseda?
Poišči besede s podobnim pomenom.
Kako pregibam to besedo?
Kako izgovorim to besedo?
V katerih frazemih nastopa ta beseda?
Kako uporabljajo to besedo v narečjih?
Od kod izvira ta beseda?
Od kdaj uporabljamo to besedo?
Kaj je kaj v slovarju?
Fran

Kaj pomeni beseda nož?

1. POMEN
RAZLAGA: priprava za rezanje iz rezila in ročaja
ZGLEDI: lupiti, rezati, strgati z nožem
zabosti z nožem

1. PODPOMEN
RAZLAGA: vojaško bodalo
ZGLEDI: nasaditi nož na puško

2. POMEN NAVADNO V MNOŽINI TEHNIKA
RAZLAGA: del stroja z rezilom, zlasti za rezanje
ZGLEDI: pritrditi nože na glavo

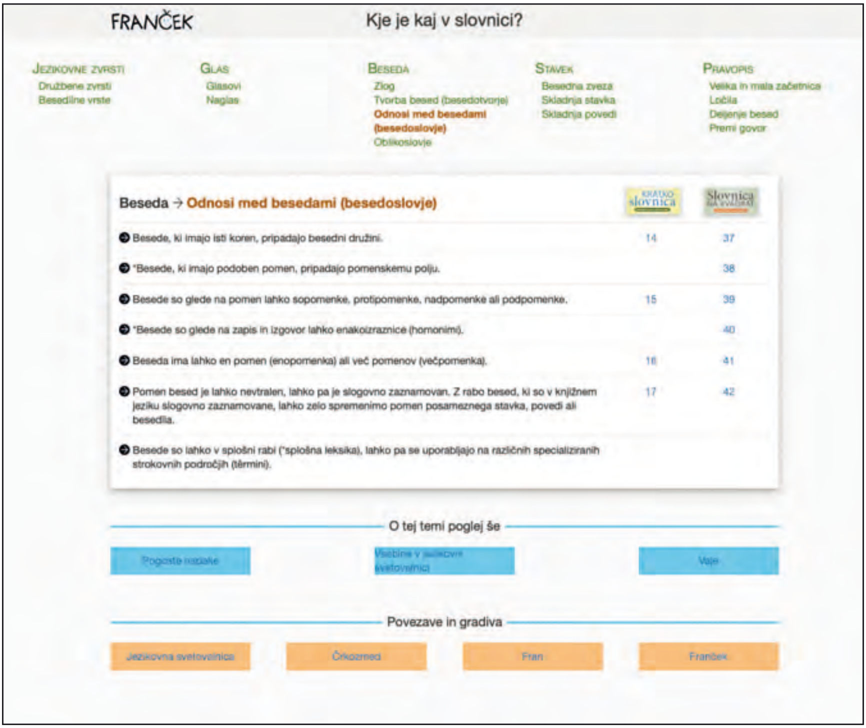
3. POMEN MNOŽINA NAREČNO
RAZLAGA: ribežen (za zelje, repo)
ZGLEDI: ribati zelje na nože

Ta opis je narejen na podlagi SSKJ2 – Slovarja slovenskega knjižnega jezika (2. izdaja), ki si ga lahko ogledaš na Franu.

Fig. 5: The Franček web portal: part of the presentation of the word *nož* ‘knife’

The Franček portal is also connected to a grammar page[14] featuring grammar topics adapted to students of primary and secondary schools (see Fig 6). The presentation of grammar is the same for all three age groups, but the explanation of grammatical topics for primary and secondary schools differs in the level of detail.

Every piece of lexicographic information is also placed within a grammatical context, which in turn forms the starting point for linking the material to the corpus of the most common errors made by schoolchildren, a special Questions and Answers language page for teachers,[15] and useful exercises.



Beseda → Odnosi med besedami (besedoslovje)	KRATKO slovnica	Slovnica
➊ Besede, ki imajo isti koren, pripadajo besedni družini.	14	37
➊ *Besede, ki imajo podoben pomen, pripadajo pomenskemu polju.		38
➊ Besede so glede na pomen lahko sopomenke, protipomenke, nadpomenke ali podpomenke.	15	39
➊ *Besede so glede na zapis in izgovor lahko enakoizraznice (homonimi).		40
➊ Beseda ima lahko en pomen (enopomenka) ali več pomenov (večpomenka).	16	41
➊ Pomen besed je lahko nevtralen, lahko pa je slogovno zaznamovan. Z rabo besed, ki so v knjižnem jeziku slogovno zaznamovane, lahko zelo spremenimo pomen posameznega stavka, povedi ali besedila.	17	42
➊ Besede so lahko v splošni rabi (*splošna leksika), lahko pa se uporabljajo na različnih specializiranih strokovnih področjih (têrmini).		

Fig. 6: The Franček web portal: a grammar page

To sum up, our services currently include: dictionaries on the Fran portal, a user-oriented language advisory service, and an introduction to dictionary use on Franček.

[14] https://kje-je-kaj-v-slovnici.franček.si/domov.

[15] https://svetovalnica.franček.si/domov.

3. Related activities

We also seek to connect all of this with sociolinguistic research. In 2016 and 2017, we thus conducted the study Slovenia's Language Policy and User Needs, which was the first of its kind to cover, with the help of a nation-wide survey, all basic sociolinguistic problems related to Slovenian (Gliha Komac 2018). Having reviewed legislative and technical documents (Gliha Komac/Kovač 2018), a group of 45 experts and researchers, who possess insight into uses, practices and needs of language users in the Republic of Slovenia and Slovenian language users outside the Republic of Slovenia due to their work and research interests, prepared a sociolinguistic description of Slovenian language community. The central part of the research project was an online survey on language use, knowledge and needs (in media, education, administration and public services, economic and social life etc.) of different groups of language users in the Republic of Slovenia and users of the Slovenian language in neighbouring countries and elsewhere in the world (Ahačič et al. 2017) in which 5,782 language users, i. e. both specialised and general, concerned by Slovenian language policy participated. This was a comprehensive attempt to actively integrate language users in the making of future language policy to the greatest possible extent. Based on this study and a number of other contributions by various researchers, the Slovenian parliament adopted the Resolution on the National Language Policy Program. The program addresses two key spheres: language education and language infrastructure (sources and technologies).

We responded immediately by compiling a new Slovenian normative guide,[16] which is based on material from the advisory pages and which cross-links rules to the dictionary section of the normative guide on the Fran web portal. Users can comment on any part of the normative guide. We will take these comments into account when preparing the final version.

4. Conclusion

Over the past eight years, the ZRC SAZU Fran Ramovš Institute of the Slovenian Language has managed to switch from a mere digital presentation of printed dictionaries to portals that are distinctly user-oriented and also provide instant feedback to lexicographers. This feedback in the form of user questions, comments, and suggestions as well as the possibility of monitoring the search statistics, has reshaped our work in practice. For example, identifying language users' problems facilitates normative work. In addition, user preferences and searches guide our selection of headwords for dictionaries for which the headwords are still incomplete (the ones that have been completed are published on an ongoing basis as part of growing dictionaries).

[16] https://www.fran.si/pravopis8.

Users have begun to see that we provide a comprehensive service. Just like on Google, where they can find the desired website, they can find all of the information on a word or multi-word unit they are interested in in one place on the Fran and Franček portals. In this way, they grow accustomed to the fact that dictionaries can help us solve various linguistic problems and, at the same time, they increasingly perceive dictionaries, corpora, and linguistic advice as direct linguistic assistance that can help them meet their daily linguistic challenges.

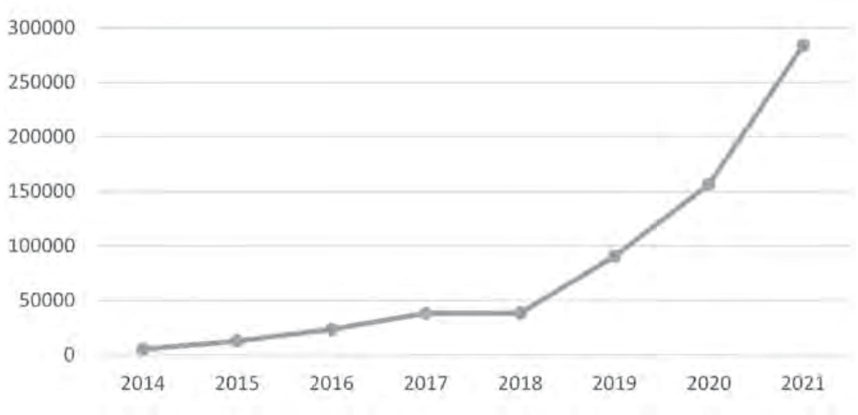


Fig. 7: The Fran web portal: average daily searches by year

References

Ahačič, K./Ježovnik, J./Ledinek, N./Perdih, A./Petric, Š./Race, D. (2021): Priprava jezikovnih podatkov za pedagoški portal o slovenščini Franček. In: *Philological Studies* 19/1, 203-224.

Ahačič, K. et al. (2017): Jeziki v Sloveniji in slovenščina zunaj nje: anketni vprašalnik. Ljubljana: ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša. https://isjfr.zrc-sazu.si/sites/default/files/anketni_vprasalnik_1_kod_0.pdf

Gliha Komac, N. (2018): Jezikovna politika Republike Slovenije in potrebe uporabnikov. In: *Slavia Centralis* 11/2, 7-15.

Gliha Komac, N./Kovač, P. (eds.) (2018): Pravna ureditev in programski dokumenti o jezikovni rabi in praksah jezikovnih uporabnikov v Republiki Sloveniji in uporabnikov slovenskega jezika v sosednjih državah in po svetu. Ljubljana: ZRC SAZU, Inštitut za slovenski jezik, Založba ZRC.

Perdih, A. (2018): Dictionary portal Fran: current state and future developments. In: *Slovanská lexikografie počátkem 21. století: sborník příspěvků z mezinárodní konference, Praha 20.-22.4.2016*, 57-65.

Perdih, A. (2020): Portal Fran: od začetkov do danes. In: *Rasprave Instituta za hrvatski jezik i jezikoslovlje* 46/2, 997-1018.

Perdih, A. (2021): Indikatorji pri homografih na portalu Franček. In: *Jezikoslovni zapiski* 27/2, 7-21.

Perdih, A./Ahačič, K./Ježovnik, J./Race, D. (2021): Building an educational language portal using existing dictionary data. In: *Jazykovedný časopis* 72/2, 568-578.

Petric, Š. (2020): Tipologija razlag v Šolskem slovarju slovenskega jezika. In: *Slavistična revija* 68/3, 391-409.

Moritz Sommet

The digital language research landscape in multilingual Switzerland

Abstract

This national report briefly discusses the current state of digital technology in language research in the Swiss context. Switzerland is known for its institutionalized multilingualism and characterized by a relatively decentralized political and administrative structure. The report presents some recent research projects in the broader field of language technology that reflect this regional and linguistic diversity and lays out current developments in digital research infrastructure that point towards increased inter-institutional cooperation and centralization. I also identify institutional key players such as regional and national research associations and institutes.

1. Introduction

Switzerland is a multilingual state with four distinct national languages and many officially bilingual or trilingual regions; its confederal political system accords relatively great weight to actors at the cantonal or communal level. Accordingly, the country has no national language institute in the sense that many other European nations do. Although established at national level on the basis of a federal law, the Swiss Research Centre on Multilingualism (RCM) depends on institutional partnerships with two universities in a bilingual canton, and it has a network of external partners from other higher education institutions in Switzerland. Its projects are often conducted in the form of partnerships, or financed by the RCM and then carried out directly by researchers at other institutions.

When trying to identify recent trends in Swiss digital language research, it should be useful to take a broad approach that is informed by both the institutional realities and the complexities of the country's linguistic and political landscape. In my effort to provide an overview of this diverse and regionally differentiated landscape, I will therefore look beyond national institutions and at interregional networks. I will first identify key institutional players both at national and regional levels. Along the way, I will present some research projects carried out by these institutions that deal with questions of digitalization and language, or that apply new forms of language technology, and that in many cases reflect the particularities of Switzerland as a multilingual nation. Finally, I will sketch out some recent developments in Switzerland's digital research infrastructure as it relates to linguistics and analyse tendencies toward stronger integration at national and supranational levels.

Language technology and digital research in linguistics are important and fast-moving fields, and any attempt at providing an overview will necessarily risk being incomplete and outdated even a few months after publication. Nevertheless, it is my hope that providing a snapshot impression of this research landscape as it exists in early 2022 will prove to be useful both for general orientation purposes and as a way to document the state of the field.

2. National and regional research institutions

Tasked by the Swiss Federation to "coordinate, introduce and conduct applied research on languages and plurilingualism",[1] the Research Centre on Multilingualism investigates various aspects of languages in the Swiss context. Sociolinguistic research by the RCM examines "how multilingualism in institutions and society impacts the political sphere, the economy and (public) administration".[2] An especially prominent part of the RCM's research deals with questions of language acquisition, the evaluation of multilingual competences, and language learning and teaching in multilingual settings. Digital language technology plays an increasingly important role in this context, as evidenced, for example, by the debate on distance learning that has acquired a particular urgency due to the recent COVID-19 pandemic. The RCM's most recent research programme, running from 2021 to 2024 and developed after consultations with the Swiss Federal Office of Culture, reflects this current shift towards digitalization.[3] Among the ten new research projects, several deal either directly or indirectly with questions raised by the pandemic. *Multilingualism in a health crisis*, for example, takes a sociolinguistic perspective on challenges in communicating COVID-19-related information, "be it information about the current situation, health issues and distancing rules, or to explain work-related rights and obligations, access to emergency financial aid, and even educational matters". The project examines channels of multilingual communication by governmental, institutional, and private organizations, such as websites, and analyses to what extent they took the needs of language minorities into account, including both speakers of official languages and residents who speak none of these languages nor English. The results should help Switzerland "to optimise the ways in which the tools of crisis management reflect the ideals of social inclusion and language sensitivity". Other projects are in the areas of

[1] Cited from the official English translation of article 18 of the Federal Act on the National Languages and Understanding between the Linguistic Communities (LangA), which was passed by the Federal Assembly on October 5, 2007 and last revised in February 2021. https://www.fedlex.admin.ch/eli/cc/2009/821/en (last access: 18-02-2022).

[2] https://centre-plurilinguisme.ch/en/about-us (last access: 18-02-2022).

[3] Cf. https://institut-plurilinguisme.ch/en/research for more detailed information about the RCM's current projects. The following citations are taken from project descriptions on the website (last access: 04-03-2022).

Computer Assisted Language Learning (CALL) and Mobile Assisted Language Learning (MALL). *Digital technology and vocabulary learning in vocational education* examines how digital learning environments can be used to support vocabulary training in schools for business, management, and services that require students to learn a national language and English. The subproject *Digital translation tools in foreign language teaching and learning*, meanwhile, looks at how tools like DeepL or Google Translate can be made useful for language teaching. The *Swiss Learner Corpus* (SWIKO), an ongoing long-term umbrella project first launched in 2016, collects text productions from foreign language learners all over Switzerland, with a view to identifying how such a corpus can be made useful in foreign language teaching and multilingual education. In the RCM's current research programme, two new subprojects will expand the scope of the database and explore further applications of corpus data in language education, adding, for instance, authentic spoken language recordings and producing teaching material from these data. The corpus, which is already being used in teaching and research at Fribourg University, will be made accessible to a wider range of external researchers. The RCM also operates the Web Portal on Multilingualism, one of the most comprehensive electronic resources for research on languages in Switzerland.[4]

As mentioned at the outset, the RCM, with its numerous partnerships with institutions of higher education, is just one element in the mosaic of Switzerland's linguistic research system. Most Swiss research on languages is conducted at the country's twelve full universities. There are also many universities of applied sciences or institutes of teacher education that conduct research on language policy and language teaching. It is therefore not possible to provide a complete overview here; instead, a few institutions and some research projects conducted by them over the past ten to fifteen years will be highlighted as examples to illustrate the range of Swiss digital research on languages and linguistics.

The University of Zurich, the largest institution of higher education in Switzerland with approximately 28,000 students, has been conducting strong research in this area for some time. Researchers at its Department of Computational Linguistics publish on a wide range of topics, such as computational text or speech processing, forensic phonetics, experimentational computer linguistics, and computational neuroscience. Many of the research projects conducted or supported by the Department apply modern language technology to the Swiss context and focus on either the country's multilingual nature or the diglossic situation in German-speaking Switzerland, i.e. the parallel use of Standard German and a variety of Swiss German dialect.[5] Various corpus linguistic projects from the last two

[4] https://centre-plurilinguisme.ch/en/centre-de-documentation#anchor13 (last access: 09-03-2022).

[5] See Studler (2012) for an examination of diglossia in German-speaking Switzerland.

decades may be cited as illustrations: *NOAH's corpus* from 2018, for instance, examines texts written in Swiss German and provides POS tagging for this language. It applies natural language processing techniques to Swiss German dialects, which are usually associated with spoken language and for which there are no standard spelling rules, a fact that made this project particularly challenging.[6] Other corpus linguistics projects made use of Switzerland's tradition of multilingual publishing to create parallel corpora. The *Text+Berg*, or 'Text & Mountain,' digital project (2008-) covers two publication series by the Swiss Alpine Club. These series have been published continuously in French, Italian, and German since 1864. They are being digitally recorded and corpus-linguistically processed. As the project description notes,

> The computational linguistic interest in the corpus lies on the one hand in the preparation of the corpus itself (automatic word type recognition, proper names/ place name recognition etc.), but also in the analysis of the linguistic data, for example to refine language models. Since the publications contain not only German, but also French and Italian texts, it makes sense to create a "comparable corpus" for multilingual questions.[7]

Similar parallel corpora created by the Department include a corpus compiled from the archives of the bulletin of the Swiss bank Crédit Suisse.[8] This magazine has been published in several languages, not just in the national languages of German, French, and Italian, but also in English and even Spanish. As with *Text+Berg*, the *Credit Suisse Bulletin Corpus* is of interest to researchers working at the interface of the digital humanities and contrastive linguistics, all the while offering a potential for researchers working with sociolinguistic approaches to discourse analysis.

The University of Zurich's Department of Computational Linguistics is but one notable Swiss institution of higher education working on such topics. The *Swiss SMS Corpus* project, compiled between 2009 and 2010 by the Department of Romance Linguistics at the same University, consists of close to 26,000 mobile text messages which were sent in by the Swiss public: 41% of the messages are in the Swiss German dialect, 28% in non-dialectal German, 18% in French, 6% in Italian, and 4% in Romansh.[9] Other noteworthy research departments in this context include the ILC Institute of Language Competence at Zurich's ZHAW School of Applied Linguistics, whose research focuses on areas such as digital linguistics and human-machine communication.[10] ZHAW's School of Engineer-

[6] https://noe-eva.github.io/NOAH-Corpus/ (last access: 09-03-2022).

[7] https://textberg.ch/ (last access: 09-03-2022).

[8] https://pub.cl.uzh.ch/projects/b4c/ (last access: 09-03-2022).

[9] https://sms.linguistik.uzh.ch/ (last access: 09-03-2022); cf. Dürscheid/Stark (2011). A more recent project in the same vein examines WhatsApp messages (Ueberwasser/Stark 2017).

[10] https://www.zhaw.ch/de/linguistik/institute-zentren/ilc/ [last access: 09-03-2022).

ing Centre for Artificial Intelligence works on the natural language processing of Swiss German dialects[11] while its Digital Discourse Lab has published *The Swiss Corpus for Applied Linguistics* (Swiss-AL), a "linguistically processed, multilingual collection of texts from key stakeholders in the field of Swiss public communication".[12] Interdisciplinary centres and groups at other universities pursue similar research interests. The University of Neuchâtel's Centre de Linguistique de Corpus (CLC) may be cited as one example.[13] OFROM (*le corpus Oral de Français de Suisse Romande*), the first spoken language corpus that consists exclusively of speech from French-speaking Switzerland, illustrates the university's activities in this field.[14] The University of Geneva's Computational Learning and Computational Linguistics research group, meanwhile, "is concerned with interdisciplinary research combining linguistic modelling with machine learning techniques",[15] while the University of Basel's Digital Humanities Lab boasts a "fast growing research agenda in digital editions, digital photography, computational linguistics and literary studies, digital reading studies and digital infrastructures".[16]

It is perhaps no coincidence that Swiss university departments frequently address the country's multilingual nature in their language related research. For speakers of Switzerland's official minority languages, digital research is not only a way of exploring the linguistic, social, and psychological dynamics of multilingual exchange but can also be a way to preserve their language and their cultural heritage. About 0.5% of the Swiss population speaks Romansh as their main language, with decreasing numbers observed from 1970 to 2020.[17] Applied research projects such as *Translaturia* (University of Applied Sciences of the Grisons) are attempting to counter the language's lack of a media presence and give it greater visibility. The project seeks to create a translation tool and develop recommendations to help companies with translation activities to better digitalize and partially automate their existing processes when using Romansh.[18] *Capeschas*, developed by the University of Teacher Education of Grisons in collaboration with the RCM, is an interactive online tool that helps with the acquisition of receptive Romansh

[11] https://www.zhaw.ch/en/research/research-database/project-detailview/projektid/5059/ (last access: 09-03-2022).

[12] https://www.zhaw.ch/en/linguistics/research/swiss-al/ [last access: 09-03-2022).

[13] http://www.unine.ch/clc (last access: 09-03-2022).

[14] http://www11.unine.ch/ (last access: 09-03-2022).

[15] https://clcl.unige.ch/ (last access: 09-03-2022).

[16] https://dhlab.philhist.unibas.ch/en/ (last access: 09-03-2022). Situated at the crossroads of digital linguistics and the liberal arts, the digital humanities have become a recent focus of interest at many other Swiss universities, including the University of Bern and Lausanne's EPFL.

[17] See the Federal Statistical Office's website: https://www.bfs.admin.ch/bfs/en/home/statistics/population/languages-religions/languages.html (last access: 11-03-2022).

[18] https://translaturia.fhgr.ch/ (last access: 11-03-2022).

skills.[19] Other projects are situated in the field of digital philology. The *Dicziunari Rumantsch Grischun* makes the vocabulary of the Romansh language accessible and has been making various digital contents available online since 2007.[20] The digital Rhaeto-Romanic Chrestomathy developed by researchers at the University of Cologne, Germany, is also worth mentioning in this context.[21]

Research is also being carried out at establishments that are attached to universities but function on an inter-institutional basis and frequently take on outside funding. In French-speaking Martigny, for instance, the federally funded EPFL University and the University of Geneva co-finance the Idiap Research Institute[22] – a non-profit research foundation that also receives funding from the local and cantonal authorities as well as from Swisscom, the country's largest telecom company. This is worth mentioning because there are generally few well known or successful commercial companies in Switzerland that deal with language technology in the commercial sector. While the country does have a start-up culture, there are, as of yet, relatively few companies that stand out in fields such as machine translation or language processing (Rehm/Uszkoreit 2012, 31). However, as the example of the Idiap Research Institute shows, technology transfer is still significant in the context of Swiss research institutions. While Idiap conducts basic and applied research in all fields related to artificial intelligence, an important part of its activities concerns linguistics and language technology: the Institute has research groups dedicated to language and cognition, natural language understanding, signal processing for communication, and speech and audio processing.

3. Switzerland's digital research infrastructure

Taking a step back from individual institutional actors and individual research projects, we can see that the digital infrastructure that supports linguistic research in Switzerland has been developing at a fast pace in recent years. As research activities are largely devolved to cantonal level with organizational and financial support from national academic associations,[23] this development has not seen much direct input from the federal administration. In Switzerland's federal digitaliza-

[19] http://chapeschas.ch/app.php (last access: 11-03-2022).

[20] https://www.drg.ch/ (last access: 11-03-2022).

[21] http://www.crestomazia.ch/ (last access: 11-03-2022). See also Neuefeind/Rolshoven/Steeg (2011).

[22] https://www.idiap.ch/en (last access: 16-03-2022).

[23] See, for instance, the significant P-5 funding programme set up by the Swiss Universities Rectors' Conference, which aims at "improving the supply of digital scientific content and creating optimised tools for processing it": https://www.swissuniversities.ch/en/topics/digitalisation/p-5-scientific-information (last access: 16-03-2022).

tion strategy, language(s) and language technology are not mentioned directly.[24] Nevertheless, the general trend is towards a greater centralization of platforms and services at national level, but in a way that takes the linguistic diversity of the country and its regional sensibilities into account. Three examples may serve to illustrate this trend.

The first example is the SLSP – the Swiss Library Service Platform (cf. Marty/Küssow 2021). Originally founded in 2015 by 15 academic institutions, the network now gathers scientific information from 475 libraries throughout Switzerland. The country used to have separate library service networks for each of its language regions. Earlier attempts to build an integrated national union catalogue were unsuccessful. The SLSP finally succeeded in establishing such a catalogue with Swisscovery, which launched in late 2020.[25] Swisscovery offers a multilingual interface in French, German, Italian, and English. The SLSP is also a consortium that acquires licences for databases and other electronic resources for university libraries. The underlying library network allows affordable interlibrary loans between the country's four language regions.

Questions of archiving and managing research data have also become of increasing relevance to Swiss language research. As Switzerland transitions to the Open Data paradigm and tries to make as many of its research data publicly available according to the FAIR principles, new technologies and platforms are emerging.[26] At the universities of Geneva and Fribourg, OLOS[27] is already in use, a transdisciplinary platform developed by the Data Life-Cycle Management (DLCM) project (cf. Burgi/Makhlouf Shabou 2021). There are also more specialized platforms of interest to linguists and language researchers, such as DaSCH (Data and Service Center for the Humanities),[28] or several solutions proposed by SwissUBase.[29] SwissUBase already serves as the basis for FORSbase, a new version of the FORS database already in use for some years among social scientists.[30] A solution specifically aimed at language researchers is currently under develop-

[24] https://www.digitaldialog.swiss/en/actionplan (last access: 16-03-2022). The Swiss Academy of Engineering Sciences' 2019 report on artificial intelligence technology mentions the potential of language technology for the sciences (Schweizerische Akademie der Technischen Wissenschaften 2019, 2).

[25] https://swisscovery.slsp.ch/ (last access: 16-03-2022).

[26] For some general impressions of recent developments, see Burgi/Echernier (2020), or the website of the swissuniversities Open Science programme: https://www.swissuniversities.ch/en/topics/digitalisation/open-science-2021-2024 (last access: 16-03-2022).

[27] https://olos.swiss/ (last access: 16-03-2022).

[28] https://www.dasch.swiss/ (last access: 16-03-2022).

[29] https://www.swissubase.ch/en/ (last access: 16-03-2022).

[30] https://forsbase.unil.ch/ (last access: 16-03-2022).

ment.[31] All these platforms are offered at national level and will likely soon replace regional or institutional repositories for research data that have already existed for some time, such as local solutions employed by the Research Centre on Multilingualism.

As a third, recent example, we can cite LiRI, standing for Linguistic Research Infrastructure. The LiRI laboratory at the University of Zurich was formally inaugurated in autumn 2021. LiRI is a project that has been in development since 2017 and aims to enable "internationally significant research in linguistics, putting Switzerland at the forefront of experimental and Big Data based research".[32] LiRI essentially consists of two services aimed at facilitating experimental and data-based language research: the first is a physical laboratory with state-of-the-art language technology used in psycholinguistic and neurolinguistics research that ranges from eye-tracking devices to machines measuring the auditory brainstem response. Additionally, LiRI offers data services such as the creation of databases to store and search data collections, or access to highly specialized software for linguistic data transcription or annotation. Against payment of a fee, these services are available to all researchers in Switzerland. *Swissdox@LiRI* is a service of particular interest to researchers working on automated approaches to mass media discourse analysis. As of early 2022, "[t]he database includes about 29 million media articles (press, online) from a wide range of Swiss media sources covering many decades, and is updated daily with about 5,000 to 6,000 new articles from the German and French speaking parts of Switzerland".[33] While additional media sources from the Italian and Romansh speaking regions of the country would undoubtedly be a highly welcome addition to Swissdox, it should be noted that the project is already continuously being expanded with improvements to the interface and data enrichment features such as POS tagging. Visualization functions, which are becoming an increasingly important tool for analysing text corpora (cf. Bubenhofer et al. 2019), will also be implemented at some point.

As all three examples show, there is an increasing tendency towards national integration in Swiss research infrastructure. This tendency is typically being supported by select university-based institutional actors, while being financed through a combination of initial federal funding and support from consortial networks that are grounded in the scientific community and receive long-term funding from various cantonal universities. Even though individual research institutions remain linguistically diverse and geographically dispersed, there is now an increasing number of offers available at national level, aimed at Swiss researchers from all parts of the country and all language regions.

[31] https://www.ub.uzh.ch/de/wissenschaftlich-arbeiten/mit-daten-arbeiten/swissubase.html (last access: 16-03-2022).

[32] https://www.liri.uzh.ch/ (last access: 16-03-2022).

[33] https://www.liri.uzh.ch/en/services/swissdox.html (last access: 16-03-2022).

4. From national to international integration

This tendency towards greater integration on a national level goes hand in hand with ongoing efforts to connect the Swiss research community with its European and international counterparts. Long-standing political frictions concerning the general framework for cooperation between the EU and Switzerland have led to disruptions in this area in recent years; as of 2022, Switzerland has been reduced to the status of a non-associated third country in the EU Framework Programme for Research and Innovation *Horizon Europe* and other related initiatives.[34] Nevertheless, Swiss institutions in the domain of language research have recently made some progress in connecting with their European partners.

Joining CLARIN (Common Language Resources and Technology Infrastructure), the principal European research infrastructure project in digital linguistics, has long been outside the scope of Swiss linguistics programmes.[35] In December 2020, CLARIN-CH was founded as the Swiss node of CLARIN Europe. Among its founding members are the universities of Bern, Lausanne, Lugano, Neuchâtel, and Zurich, as well as the University of Applied Sciences Zurich, and the Swiss Academy for the Humanities and Social Sciences (SAGW).[36] CLARIN-CH is now recruiting additional interested parties, touting benefits such as "[i]ncreased international visibility for Swiss corpus-based projects, corpora and linguistic databases, tools and infrastructure" or access to European infrastructure programs, services, and funding opportunities.[37] Among the association's dissemination efforts is a 'Tour de Suisse' with information sessions at all research and academic institutions throughout the country that are interested in gaining access to CLARIN resources and infrastructure. The national and scientific coordination of the CLARIN-CH network is being assumed by Zurich-based researchers, and CLARIN-CH coordinates with the above-mentioned LiRI project, which is also based at the University of Zurich. The network describes its principal mission as follows:

1) Obtain Switzerland's CLARIN membership and give Swiss researchers access to the entire CLARIN infrastructure.
2) Bring together the Swiss community using language resources and create national working groups.
3) Foster the sharing of expertise and of resources.
4) Encourage the initiation of national and international collaborations.[38]

[34] Cf. the Confederation's official website on this matter: https://www.horizon-europe.ch (last access: 21-03-2022).

[35] Cf. Eskevich et al. (2020) for an introduction to CLARIN and the official website for up-to-date information: https://www.clarin.eu (last access: 21-03-2022).

[36] https://clarin-ch.linguistik.uzh.ch/ (last access: 21-03-2022).

[37] https://clarin-ch.linguistik.uzh.ch/_media/poster_clarin-ch_september10.pdf (last access: 21-03-2022).

[38] https://clarin-ch.linguistik.uzh.ch/_media/vals-asla_forum_clarin-ch_february2022.pdf (last access: 21-03-2022).

In pursuit of this mission, CLARIN-CH is preparing an application for observer status at CLARIN-EU.

A similar status has already been achieved by the Swiss consortium for DARIAH, the Digital Research Infrastructure for the Arts and Humanities. Established in 2014 as a European Research Infrastructure Consortium, DARIAH-EU describes itself as “a network of people, expertise, information, knowledge, content, methods, tools and technologies from its member countries“ that “aims to enhance and support digitally-enabled research and teaching across the arts and humanities”.[39] Its Swiss partner DARIAH-CH was founded in 2018.[40] Coordinated by the Basel-based DaSCH (see above) and with significant initial support from researchers at the University of Neuchâtel, DARIAH-CH aims to connect Swiss institutions with European digitally-enabled research in the arts and humanities or the teaching of digital research methods. Among its members are the Swiss Academy of Humanities and Social Sciences and the universities of Basel, Bern, Geneva, Lausanne, Neuchâtel, and Zurich, as well as Lausanne’s EPFL. While DARIAH-CH is less directly invested in linguistic research in the narrow sense of the term than CLARIN-CH, the network’s activities in the areas of Open Linked Data or digital text analysis and curation should be of interest to Swiss researchers active in digital linguistics and adjacent fields.

5. Conclusion

Switzerland has a diverse and regionally differentiated research landscape in the field of linguistics, language teaching, and language planning. While some of the institutions introduced in this paper stand out in terms of their research output and the resources they dispose of, this generally also holds true for digital linguistics. The research projects I have briefly presented here do not just reflect this institutional diversity but also the ongoing necessity to consider the country’s linguistic diversity as well as the perspectives and needs of its language minorities. They also demonstrate that the generally popular notion of ‘multilingualism as a resource’, which has occasionally invited criticism (cf. Duchêne 2011), holds some weight in Switzerland — at least when it comes to acquiring and exploiting material and sources for linguistic research.

The general tendency towards greater national and international integration noted in this paper will likely increase in the years to come. This integration process is supported from within the scientific community and driven by cooperative efforts between actors at regional and national levels. We could, however, also think of it as almost a function of the global nature of the digitalization process

[39] https://www.dariah.eu/about/dariah-in-nutshell/ (last access: 21-03-2022).

[40] https://dariah.ch/ (last access: 21-03-2022).

per se, which both encourages and enables processes of organizational centralization and standardization. It remains to be seen how these developments will co-exist with the confederate system and the multilingual reality in Switzerland.

References

Bubenhofer, N./Rothenhäusler, K./Affolter, K./Pajovic, D. (2019): The linguistic construction of world: An example of visual analysis and methodological challenges. In: Scholz, R. (ed.): *Quantifying approaches to discourse for social scientists*. Cham, 251-284. https://doi.org/10.1007/978-3-319-97370-8_9.

Burgi, P. Y./Echernier, L. (2020): A review of the Swiss Research Data Day 2020 (SRDD2020): 48 experts shared their experiences on emergent approaches in open science. In: *Revue Électronique Suisse de Science de l'information* 21. http://www.ressi.ch/num21/article187.

Burgi, P. Y./Makhlouf Shabou, B. (2021): Le projet Data Life-Cycle Management (DLCM) en Suisse: une gestion des données de la recherche pensée pour ses utilisateurs. In: *I2D – Information, donnees documents* 2 (2), 87-95.

Duchêne, A. (2011) Neoliberalism, social inequalities, and multilingualism: The exploitation of linguistic resources and speakers. In: *Langage et societé* 136, 2 (6 July 2011), 81-108. https://doi.org/10.3917/ls.136.0081.

Dürscheid, C./Stark, E. (2011): Sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland. In: Thurlow, C./Mroczek, K. (eds.): *Digital discourse: Language in the new media*. Oxford, 299-320. https://doi.org/10.1093/acprof:oso/9780199795437.003.0014

Eskevich, M./de Jong, F./König, A./Fišer, D./van Uytvanck, D./Heuvel, H. (2020): CLARIN – Distributed language resources and technology in a European infrastructure. In: Rehm, G./Bontchev, K./Choukri, K./Hajič, J./Piperidis, S./Vasiļjevs, A. (eds.): *Proceedings of the 1st International Workshop on Language Technology Platforms* (*IWLTP 2020*). Marseille, 28-34. https://arne.chark.eu/anthology/2020.iwltp-1.5/.

Marty, T./Küssow, J. (2021): SLSP and Swisscovery – a Swiss Success Story on How to Create a Multilingual, National Library Network. In: *ABI Technik* 41, 2, 64–70. https://doi.org/10.1515/abitech-2021-0015.

Neuefeind, C./Rolshoven, J./Steeg, F. (2011): Die Digitale Rätoromanische Chrestomathie-Werkzeuge und Verfahren für die Korpuserstellung durch kollaborative Volltexterschließung'. In: *Multilingual Resources and Multilingual Applications: Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology* (*GSCL 2011*). Hamburg, 163-168.

Rehm, G./Uszkoreit, H. (2012): Sprachtechnologie für das Deutsche. In: Rehm, G./Uszkoreit, H. (eds.): *The German language in the digital age*. Heidelberg, 17-37.

Schweizerische Akademie der Technischen Wissenschaften (2019): *Künstliche Intelligenz in Wissenschaft und Forschung*. https://www.sbfi.admin.ch/sbfi/en/home/eri-policy/eri-21-24/cross-cutting-themes/digitalisation-eri/artificial-intelligence.html.

Studler, R. (2017): Diglossia and bilingualism. High German in German-speaking Switzerland from a folk linguistic perspective. In: *Revue Transatlantique d'études Suisses* 6, 7, 39-57.

Ueberwasser, S./Stark, E. (2017): What's up, Switzerland? A corpus-based research project in a multilingual country. In: *Linguistik Online* 84, 5. https://doi.org/10.13092/lo.84.3849.

EFNIL project report

Sabine Kirchmeier/Aino Piehl/Johan Van Hoorde/Júlia Choleva/
Katrin Hallik/Cecilia Robustelli

ELIPS – European Languages and their Intelligibility in the Public Sphere

Abstract

This article contains an analysis of the data survey ELIPS (https://elips.efnil.nytud.hu/). ELIPS is the acronym for *European Languages and their Intelligibility in the Public Sphere*, one of EFNIL's major projects. The project focuses on the use of the official languages of various European countries as instruments for legislation, government and public administration. Attention was paid, amongst others, to the use of plain and easy-to-read language, the availability of high-quality terminology for legislation and public administration, the existence of practices and policies regarding diversity in society (linguistic and cultural minorities, gender diversity). The data survey also focuses on the training facilities in these domains for civil servants and on national participation in international, collaborative structures. The survey was conducted by EFNIL in 2018-19 and contains information from 24 European countries covering 27 languages.

The article starts with a short description of the various subdomains covered by the survey as well as the issues and trends at stake within each of them. This forms the basis for a detailed presentation of the data, with a series of tables and figures that will enable readers to gain a good overview of the situation in Europe and to compare countries. The article ends with a series of recommendations, both general ones for stakeholders active in these fields and specific ones for EFNIL as a collaborative network of national language institutions.

1. Introduction to ELIPS

ELIPS is a project organised by EFNIL, the European Federation of National Institutions for Language. The acronym ELIPS refers to (the use) of ***E****uropean* ***L****anguages and their* ***I****ntelligibility in the* ***P****ublic* ***S****phere*, which underlines the aims of the project, namely to examine the use of European languages as instruments of communication for government, legislation and public administration and to find ways to promote interest in ensuring good quality communication by authorities.

As EFNIL's mission is to gather and publish information about language use and language policies within Europe, it is natural that language use by public authorities falls within the scope of these activities. When ELIPS was initiated in 2017, it was decided that its first action would be a survey in order to map the situation regarding language use by public authorities in the countries and language areas that are represented within EFNIL. The questionnaire was sent to member institutions in 2018-2019. The results were analysed in 2020-2021 and a special ELIPS website was created to present them.

The ELIPS survey is a pioneer in mapping Europe-wide the engagement of public authorities in domains important to communication with citizens. Earlier, plain language requirements placed on authorities and their activities in implementing those requirements were only examined in an international but limited pilot survey carried out by the *Plain Language Association International* and the Portuguese plain language organisation *Claro*. This 2017 survey included New Zealand, Portugal and the United States and also covered the opinion of citizens on the quality of the authorities' communications (Miguel Martinho 2017).

National surveys have partly covered the same topics. For example in Sweden the plain language activities of the authorities have been surveyed regularly since 1994. In Finland the plain language work of central government agencies and municipalities were investigated in a series of surveys in 2012-2017 and the comprehensibility of Finnish language versions of EU legislation was studied in 1998, 2006 and 2018. In Estonia, a survey was conducted in 2021 by the local plain language community to gather information and best practices of plain language in operation in various public authorities (Cf. Hansson 2020, Piehl 2019, Viertiö 2011).

The information collected through the ELIPS survey is meant to serve as a reference base for further activities within the project, e.g. for proposals, conferences and partnerships. Although the focus of ELIPS remains within the domain of the relationship between language and society, it widens the scope of EFNIL's activities from monitoring and promoting the status of national and minority languages to promoting the *quality* of communication by authorities.

Information about the ELIPS project is available on the web pages of the project at http://www.efnil.org/projects/elips. This gives each member institution of EFNIL and, indeed, everyone interested in these issues the chance to compare their national situation with other language areas and member states represented within EFNIL. Acquiring information about the actors and activities in play will hopefully serve, in turn, as a basis for further development, e.g. for formulating policies and strategies or searching for partner organisations for projects of common interest. It would also be desirable for the survey to inspire more academic research on its topics so as to provide a basis for development efforts.

ELIPS covers the following topics:

- plain language policies and actions;
- easy-to-read language policies and actions;
- terminology policies and actions;
- policies and actions on the use of other languages, gender, cultural and sexual diversity;
- training of information providers in public institutions;
- collaboration between the translation services of EU institutions and experts in member states.

2. Domains examined in the ELIPS survey

It is increasingly recognised that the language used by authorities has a fundamental impact on the functioning of society: the comprehensibility of authorities' communications affects citizens' access to rights, their legal protection and, finally, their trust in society. Good communication makes it possible to participate in and influence the development of society and to interact with authorities. An important aspect is that good communication helps the administration to function efficiently.

The ELIPS survey examines different aspects of the language used by public authorities. All those aspects, i.e. policies and practices for plain language, easy-to-read language, terminology work, taking account of societal diversity, training public officials and collaborating with EU linguistic services, contribute to successful communication and the smooth functioning of authorities.

2.1 Plain Language

Worldwide interest in the comprehensibility of the language used by public authorities resulted in plain language movements being launched in several countries. The topic had been discussed now and again before, but in the 1970s authorities started to respond on a larger scale to calls for clearer communication (see e.g. Ehrenberg-Sundin/Sundin 2015; Piehl 2008; Schriver 2017). It can be considered a necessary (albeit not sufficient) condition for an effective democracy, allowing citizens to exercise their rights and participate in the management of common issues. Thus the growing demands for plain language were connected to other movements demanding a more democratic and equal society in the 1960s and 1970s.

At first, the focus was on the complexity of sentence structure and difficult words used in communications with citizens but over the following four decades the field evolved to include coherence, text structure, tone of voice issues and information design as well as accessibility and the demands of originally or increasingly multicultural societies. Thus the focus has shifted from readability towards usability and, from there, towards the legitimacy of the government; likewise it has shifted from the text itself to the process and conditions of its creation (Ehrenberg-Sundin/Sundin 2008, 269-277; Schriver 2017: 343, 345, Tiililä 2018).

An example of both understanding the need for trust and the impact of circumstances on the success of a plain language policy is the Estonian plain language project that came into life in March 2020. Within a few days the Estonian government created a web page to inform people about the new regulations and restrictions related to the COVID crisis. Information from various government agencies dispersed over several websites was assembled on one platform and it urgently needed structure and good linguistic assistance.

A team of volunteer Estonian language editors and Russian and English translators was compiled to assist the government with the platform. For the plain language activists, this was a great opportunity to get a hands-on introduction to government communication and to train the editors and translators on the basics of plain language. Plain language guidelines were sent to officials composing the original texts in government agencies. The volunteer project lasted four months until the situation calmed down.

This example of volunteer enthusiasm linked to a government's need in a social crisis shows that efficient solutions can be created in a short time and with scarce resources. Plain language guidelines, text structures and terminology will remain in the text corpus of the government and will keep creating change.

2.1.1 Concept of plain language

In the questionnaire for the ELIPS survey, plain language is described for the respondents as follows:

> By plain language we understand any communication that uses wording, language, grammatical structures and information design aimed at making meaning as clear and therefore as effective as possible in order to offer its audience the best possible chance (a) of understanding it immediately and (b) of readily finding in it what it needs or expects, (c) of using the information it contains and/or (d) of performing the actions that are required.

This closely resembles the definition developed by the International Plain Language Federation (it should be kept in mind that the term plain language also refers to communications by businesses and NGOs):

> *A communication is in plain language if its wording, structure, and design are so clear that the intended readers can easily find what they need, understand what they find, and use that information.*[1]

This definition by the International Plain Language Federation has existed since 2010 (see Cheek 2010). It was developed jointly by plain language organisations that are members of the Federation (see Section 2.1.2). Before choosing this type of definition, possible approaches were discussed on the basis of existing definitions. The options were a numerical, formulae-based definition (e.g. readability tests), an element-based definition (focusing on linguistic and visual features) and an outcome-based definition (focusing on readers' ability to use the texts).

Existing definitions do not represent any of these types in a pure form but combine characteristics of two or all three types. Examples of element-based definitions are found, for example, in Finnish (2003) and Swedish (2008) legislation.

[1] See definitions on the website of the International Plain Language Federation: https://www.iplfederation.org/plain-language.

The Finnish law requires that public authorities use appropriate, clear and comprehensible language (*asiallista*, *selkeää ja ymmärrettävää*) while the Swedish Language act requires that it is cultivated, simple and comprehensible (*vårdat*, *enkelt och begripligt*).

The outcome-based type of definition was chosen by the Federation because readers' benefits and reading experience have become crucial in plain language work. The definition is intended to apply regardless of the language and the medium. It allows for flexibility since different audiences and media have different needs. Numerical and element-based approaches have by no means been discarded; they are used to support the approach which is based primarily on outcomes (Cheek 2010, 9). Based on this definition, an ISO standard for plain language is currently being developed in a working group that has experts from 25 countries.

It is worth remembering that *plain language* is not the only English expression that refers to the concept of comprehensible, functional or effective language, although it is in the process of becoming the most commonly used. *Clear language*, *clarity*, *comprehensibility* and *intelligibility* are also used to refer to the same concept. The term plain language has been criticised for creating a false image by linking the concept mentally to something simple and childish. This does not correspond to the purpose of plain language, however, since the aim is not to simplify the content but to ensure clear, comprehensible expression of meaning and usable communications by administrations and the judiciary, also in text types which are not only addressed at lay persons (see, for example, Kimble 2016.) It should be noted that the terms *easy-to-read language* (see Section 2.2) and plain language refer to two different concepts.

Preferring the image of clarity to that of simplicity has probably had a bearing on the choice of the equivalent term in several languages. For example, the following languages rely on *clear*: *klarsprog* (Danish), *selge keel* (Estonian), *selkeä kieli* (Finnish), *klarspråk* (Norwegian, Swedish), *linguagem clara* (Portuguese) and *lenguaje claro* (Spanish) while German and Romanian prefer the image of *plainness/simplicity*, e.g. *einfache Sprache* (German) and *limbaj simplu* (Romanian). Greek uses both terms related to *clarity*, i.e. *σαφής γλώσσα*, and to *plainness*, i.e. *απλή γλώσσα*, the latter being the one which seems to be most commonly used. It should be kept in mind that any term equivalent to *plain language* has not yet established itself in many languages and that expressions equivalent to *comprehensible language* are also common.

2.1.2 Two international organisations Clarity and PLAIN

International cooperation between actors promoting plain language seems to have gained momentum especially since the 1990s, when it was facilitated by easier contacts to other countries provided by the Internet and email. The plain language community cooperates on many levels, sharing expertise and advocating the use of plain legal language instead of legalese. Worldwide there are two big international organisations, in addition to many local plain language organisations that have been set up by plain language activists.

Clarity[2] is the oldest and largest international plain language organisation, founded in 1983, with more than 650 members in 50 countries and official representatives in around 30 countries. Its members are plain language practitioners – writers, editors, researchers, consultants and trainers, judges, lawyers, government officials, scholars and teachers as well as corporate and NGO representatives.

The parallel international organisation, *Plain Language Association International*[3] (PLAIN), has likewise created a support network for plain language practitioners around the world. The growing network includes members from over 30 countries working in clear communication in at least 15 languages.

The European Commission is one of the organisations working on clear writing as a way of providing better services to EU citizens. The Commission aims at improving the quality and clarity of its written communication. Its administrative bodies have been running a clear writing campaign for 10 years, encouraging their staff to put clear writing principles into practice and change the drafting culture at the Commission.

The European Union's booklet *How to Write Clearly*[4] is available in the 24 official languages of the EU.

2.1.3 Other international activities

Clarity and *PLAIN* have English as their working language; although the use of other languages is encouraged in conferences and on the websites, it occurs on a limited scale. There is clearly a need for gatherings conducted in other languages and there are a few European conferences and networks for plain language activities. For example the German Federal Ministry of Justice and Consumer Protection has organised five symposia since 2012 about comprehensibility in legal provisions where the languages used are German and English.

There are also conferences where English is not an option. The Nordic countries have organised biannual plain language conferences since 1998 where presentations are held in Danish, Norwegian or Swedish. Participants from the other Nordic countries are expected to understand and communicate in these. The Comprehensible Public Administration and Government Network in Belgium and the Netherlands (*Netwerk Begrijpelijke Overheid*) coordinates and stimulates plain language-related collaboration between organisations in the two countries in Dutch.

[2] See website https://www.clarity-international.org/.

[3] See website: https://www.iplfederation.org/plain-language.

[4] See online version: https://ec.europa.eu/info/sites/default/files/clear_writing_tips_en.pdf.

2.2 Easy-to-read language

Easy-to-read language (or easy language) is a form of language which is simplified in order to make information accessible to people with restricted reading and writing skills. The reason may be, for example, intellectual or developmental disabilities, poor competence in the official language of a country or even a temporary illness or crisis. The reading abilities of target groups for easy-to-read language vary and the level of simplification in easy-to-read texts varies accordingly. The *Swedish Agency for Accessible Media*[5] gives this description of basic types of simplification:

> What distinguishes easy-to-read books is, among other things, that they are written with easy everyday words, short sentences and straightforward and simple actions. There are few lines of text on each page and the text is often supported by explanatory images.

No internationally agreed definition of easy-to-read language exists, perhaps because the understanding of who belongs to target groups of easy-to-read language varies from one European country to another. However, there are national definitions (see Lindholm and Vanhatalo 2021). For example the *Finnish Centre for Easy Language*[6] defines easy-to-read language like this:

> Easy Finnish […] is a form of Finnish where the language has been adapted so that it is easier to read and understand in terms of content, vocabulary and structure. It is targeted at people who have difficulties with reading or understanding standard language.

The equivalent terms for easy-to-read language reflect the perception of the concept as they often include the word for ‘easy’, for example *leichte Sprache* (German) or *lätt språk* (Swedish), etc.

Easy language user organisations cooperate internationally or within Europe (e.g. Inclusion Europe), as do providers of easy language services and researchers of the subject. The first international conference on easy-to-read language research was held in 2019.

Easy-to-read language and plain language (see Section 2.1) are often confused with each other. It is understandable as the concepts are close. When public authorities use both easy and plain language the aim of both is to adjust language so as to give readers a better chance to know what they are entitled or obligated to do and to take care of their business with public authorities without undue difficulties. The target groups differ and the means are partly different but together the two varieties cover much of the needs of the entire population of a country and contribute to the goals of accessibility, inclusion and empowerment of all members of a society.

[5] See website: https://www.mtm.se/var-verksamhet.

[6] See website: https://selkokeskus.fi/in-english/guidelines-and-instructions/definition-and-background/.

2.3 Terminology

It is self-evident that the use of languages as instruments of legislation, government and communication by public authorities implies the use of specific terminology. This terminology is meant to increase precision and clarity within these domains, especially for communication between domain experts. For non-experts, the use of this terminology can complicate understanding and for this reason its use is often discouraged for communication to the general public.

2.3.1 Definitions and distinctions

Terminology is used to refer to groups of specialised words and their meanings within a particular field but also to refer to the scientific study of these groups of words and concepts as well as their characteristics, use and behaviour. In this article and the data survey on which it is based, the word is used almost exclusively to refer to the first meaning. In this way we can speak about the terminology of legislation (e.g. law, decree, regulation), of public governance (e.g. legislature, motion of no confidence) or of administrative law (e.g. appeal procedure, right of refusal). There is also terminology for the sciences, like quantum mechanics or thermodynamics, and technical branches, like computing and the construction industry.

Unlike the ordinary meaningful elements in language we call words, terms have specific meanings in a particular domain and situation and normally come into being by explicit stipulation (‘*the term x in this text/domain is used to refer to y*’) in order to avoid ambiguity, polysemy and connotations that might influence the interpretation and which characterise a great deal of our ‘normal’ words. Terms are not only single words but can also be compounds and multi-word expressions.

Sometimes terminology and jargon are considered to be synonyms but quite often a distinction is made. Jargon is a broader concept than just terms and refers to the linguistic characteristics of a specific language community. It does not only consist of terms in the real sense of the word but also of all kinds of words, expressions, formulations, stylistic registers and sentence patterns etc. that help to create a specific group language as the binding element of a social entity. Thus, the goal of jargon is not (only) to facilitate precision within a field of interest but also to create a specific community, a feeling of belonging among members of the same social group, in other words a group identity. Jargon functions along the demarcation lines of inclusion – exclusion. By using a certain jargon, persons manifest themselves as members of a community. People who do not know how to communicate in that proper way will be regarded as outsiders.

Needless to say, language use within legislation, government and public administration is not only characterised by the use of specific terminologies but also contains linguistic features that may be characterised as being part of the jargon of inner crowds, be it juridical experts, political actors or civil servants. These linguistic elements are not included in this survey.

2.3.2 Terminology work in the public sphere

In many countries terminologies governing the public sphere are the object of explicit action or policies. Terms are stipulated and agreed upon, collected and described and may be the object of unification or standardisation if there appear to be too many discrepancies. The actors involved in these processes differ from country to country and may involve ministries and other public authorities, official translation services, language institutes and even institutions responsible for normalisation and standardisation.

All sorts of problems concerning terminology may arise and may become the object of explicit action, for example:

- terminological differences within a given domain and disagreements between specialists in a given domain;
- terminological differences between domains that are closely related, e.g. between the economic and social spheres of public governance;
- differences between countries where a given language is used as the instrument of legislation, government and public administration as a result of broader sociocultural differences and traditions as well as official authorities and structures between these countries. These are so-called bicentric or pluricentric languages relating, for example, to the official terminology of French-speaking Belgium, the French language community of Switzerland and France;
- differences between language varieties within the same language in one country, like between the two varieties of Norwegian and between a sophisticated administrative language and a more vernacular one in Greek;
- differences and discrepancies between different languages that are used as communicative instruments in the same country, for instance between Swedish and Finnish official terminology in Finland or between Dutch and French terminology in Belgium;
- differences and variation between the terminologies used in separate countries within a given domain and the terms used for the same domain by institutions belonging to the European Union, e.g. by the European Commission and its directorates-general.

Apart from actions which address issues concerning the collection, description and unification or standardisation of terminology, many countries are also concerned about the existence of good, acceptable terms in their own official language using native lexical elements and following proper word formation processes as alternatives to terms borrowed from other languages, in most cases from English. These policy actions focus on the production and implementation of so-called terminological neologisms. Countries with active policies in this area include France, Greece and Norway.

2.3.3 International cooperation

Terminology work is also the object of international collaboration. Almost all collaborative terminology structures are not specific to the field of public governance and administration but cover all sectors that are relevant to terminology work. There are also international exchange structures between public administration bodies. For them terminology is often only one area of collaboration among others.

The *European Association for Terminology*[7] (EAFT-AET) has more than 50 institutional member organisations from all over Europe. It promotes the professionalism of terminology work and stimulates cooperation between its member institutions. EAFT has its secretariat in Barcelona, Spain.

TermNet[8] is a global network for terminology founded on the initiative of UNESCO, with the aim of stimulating collaboration and sharing expertise. It has its secretariat in Vienna, Austria.

Another collaborative structure in the field of terminology which is also based in Vienna is *Infoterm*,[9] which promotes and supports the cooperation of existing as well as the establishment of new terminology centres and networks. The ELIPS questionnaire did not explicitly ask about membership of Infoterm.

The *Conference of Translation Services of European States*[10] (COTSOES) is a platform of exchange and collaboration between 52 translation services from 20 different countries. Collaboration and sharing best practices in the field of terminology is one of the four main areas of COTSOES for which there is a specific working group.

The institutions of the European Union are also important for terminology cooperation on a European level. There is an inter-institutional database for terminology, called *IATE*[11] (*Interactive Terminology for Europe*) involving important collaboration between terminology actors belonging to member states. On the initiative of the Directorate-General for Translation there are also collaborative structures for specific official European languages in which the EU translation services collaborate with national partners in specific language areas. Examples

[7] See website: https://www.termcat.cat/en/european-association-terminology-eaft.

[8] See website: https://www.termnet.org/.

[9] See website: http://www.infoterm.info/.

[10] See website: http://www.cotsoes.org/.

[11] See website: https://iate.europa.eu/home.

are REI (*Rete per l'eccellenza dell'italiano istituzionale* – the network for the excellence of the Italian institutional language) and the *Interinstitutionelle Terminologiegruppe Deutsch* (Interinstitutional terminology group for German).

A collaborative network and tool that deserves a special mention is the *EuroTermBank*,[12] which is the largest centralised terminology bank for languages of the European Union and Icelandic. Through its harmonisation, collection and dissemination of public terminology resources, *EuroTermBank* strongly facilitates the enhancement of public sector information and strengthens the linguistic infrastructure in new EU member countries.

The last network organisation that needs to be mentioned is *Nordterm*,[13] the association of organisations and societies in the Nordic countries which are engaged in terminology work, training and research.

2.4 Diversity

Our societies are diverse. As a result, in some way or another, legislation, government and communication by public authorities, especially between these authorities and the general public ('citizens'), have to cope with this diversity, even more so as the sensibility for diversity in society has rapidly increased over the past decades. Coping with these aspects in a proper way has increasingly become a challenge for public governance and public authorities. In many cases they also constitute a challenge for our languages themselves and the linguistic and stylistic choices that are (or are not) available to express and acknowledge this diversity.

Important diversity aspects in our society are, for instance:

- the presence of languages and language communities other than the dominant, so-called official, language of the country, including minority languages with long traditions in our societies but also languages of recent migration and non-verbal sign language;
- gender diversity, the visibility of male and female persons and increasingly also acknowledgement of a more nuanced, non-binary approach to gender identities closely related to the gender phenomenon;
- diversity of sexual preferences and identities;
- social diversity, e.g. of social classes, degrees of schooling, cultural backgrounds, religious and ideological convictions;
- physical differences such as skin colour;
- functional disabilities.

There is an increasing conviction that all communication, verbal and non-verbal, should reflect society as it really is, in all its really existent variety and variation,

[12] See website: https://www.eurotermbank.com/.

[13] See website: http://www.nordterm.net/.

in order not to exclude certain categories of citizens, not to discriminate against them or to conceal their existence. Most if not all of the aspects of diversity mentioned above are subject to discussion and even struggles within our societies, including strong opposition towards this diversification and especially towards forms of linguistic engineering in order to cope with diversity issues.

In the Flemish region of Belgium, for instance, there is a language law that forbids public authorities and their civil servants to communicate in languages other than the official language(s) of the region, even if this means that crucial information, for instance in relation to public health and concerning all kinds of social regulations, does not reach certain categories of citizens. Authorities in other countries do use other languages on certain, well-specified occasions in cases where the nature of the information and its accessibility for the population at large is considered crucial to inclusion, democracy and the active participation of citizens.

Certainly gender and sexual identities are increasingly a topic of discussion in society, with sometimes contradictory and conflicting strategies and attitudes, even among those in favour of diversity policies. In some language communities there has been a tendency to systematically distinguish between male and female, even to the point of changing word formation patterns in order to produce female designations for functions and professions which did not exist before. This diversification strategy is often considered detrimental to a more nuanced, non-binary approach to gender identity, including all identities on the LGBTQIA+ spectrum. For this reason, in other societies there is the opposite tendency towards gender neutral communication involving, for example, the introduction of a gender neutral pronoun for people instead of the binary he/she dichotomy, for instance in Swedish with the neutral third person singular pronoun *hen*.

These are only a few examples of diversity issues, the strategies at stake and discussions about them within societies. These aspects are the focus of part 4 of the ELIPS data survey, revealing that in almost all countries many of these diversity aspects have only become the object of explicit policy measures relatively recently.

2.5 Training

The training of public officials is an important factor in maintaining good governance and enabling public sector agencies to meet the requirements of a developing society. Many, if not all, domains examined by the ELIPS survey call for skills that are unlikely to have been included in the regular education and training of civil servants.

Training is often purchased as a service provided by various actors such as government research and expert institutions, universities, NGOs, enterprises or individual experts. It may be organised as in-house training or as courses offered by the providers. Increasingly, lectures and courses are held as webinars and self-learning courses are offered on digital platforms; using digital media offers flexibility in time and space. This development has been speeded up by the Covid-19 crisis.

In some countries university courses are available for those who wish to improve their competence in plain language, easy-to-read language and terminology studies. Such a plain language course in English was created at the University of Antwerp in international cooperation and partly with EU funding, although at the moment it is not available. Easy-to-read courses are part of Finnish language studies at the University of Helsinki for example. In Sweden it has been possible to complete an academic degree for plain language consultants since the 1970s.

2.6 The European Union

The influence of the EU on the language of legislation and administration in member countries is significant. The leading principle guiding the language regime of the EU is multilingualism: all legislative proposals and many other texts are translated into its 24 official languages by translators who mostly come from countries where the language they translate into is spoken. In many languages, the legal language of the EU has developed into a variety that is different enough from the national legal language to be called a eurolect (see Mori 2018). Sometimes this eurolect is regarded as more comprehensible and usable than the national variety, sometimes vice versa (cf. Mikhailov/Piehl 2018).

In order to achieve functional legislation in its official languages, EU translation units have established contacts with public officials and language experts in member states in order to consult them about various linguistic issues. These contacts may be informal, i.e. built on personal acquaintances, but there are also structured, more official networks and platforms which have often been found to be useful (Somssich et al. 2010, 46-47). Collaboration facilitated by such platforms takes various forms. There may be a need for guidance in language problems (e.g. textual, syntactic, terminological) when discussing new terminology, creating translation tools or training and interaction on other topics.

3. Participating countries and languages represented

Twenty-three out of the 34 EFNIL member institutions and one additional institute representing 24 countries and 27 official languages provided information in the ELIPS questionnaire. In total, there were 28 respondents:

- Austria
- Belgium (Flemish Community)
- Bulgaria
- Denmark

- Estonia
- Finland (answers regarding Swedish)
- Finland (answers regarding Finnish)
- Germany
- Grand Duchy of Luxembourg
- Greece
- Hungary
- Iceland
- Ireland (except Northern Ireland)
- Italy
- Latvia
- Lithuania
- Malta
- The Netherlands
- Norway
- Portugal
- Slovak Republic
- Slovenia
- Sweden
- Switzerland
- United Kingdom (England)
- United Kingdom (Wales)
- United Kingdom (Northern Ireland)
- United Kingdom (Scotland)

In some countries with more than one official language, the questionnaire was answered separately for each language. In some cases, a country has identical provisions for different languages and in some cases the provisions differ. For instance, the legal provisions in Finland for Finnish and Finland-Swedish are almost the same, whereas in the UK they differ for Welsh and English. In other cases, the same language is spoken in different countries with different provisions. Other countries, e.g. Switzerland, chose to fill in the questionnaire just once covering all official languages. For Belgium, on the other hand, there is only information regarding the Dutch-speaking part of Belgium, i.e. the Flemish Region and the Dutch-language community of the bilingual Brussels Capital Region, and the situation in the French-speaking areas of Belgium might be completely different.

Therefore, the statistical data in the survey is based on the answers provided by each respondent, not on countries or languages as a whole, e.g. there is one response from Finland for Finnish and one for Swedish although the provisions for the two languages in most cases may be identical. This should be kept in mind when interpreting the data.

4. Project group

The plan to conduct a survey as the first stage of the ELIPS project was initiated by the project group which also designed the questionnaire. The group was nominated by the executive Committee of EFNIL in 2017. In conducting the survey the group was assisted by the Danish Language Council and Sabine Kirchmeier did the main part of setting up the website and analysing the results. The group that conducted the survey consisted of the following persons:

- Aino Piehl, Finland;
- Cecilia Robustelli, Italy;
- Johan Van Hoorde, Belgium/the Netherlands;
- Júlia Choleva, Slovakia;
- Katrin Hallik, Estonia.

The following persons contributed to the work in its earlier stages: Anne Kjærgaard, Denmark; Nathalie Marchal, Belgium; Daiva Vaišnienė, Lithuania.

5. The ELIPS survey

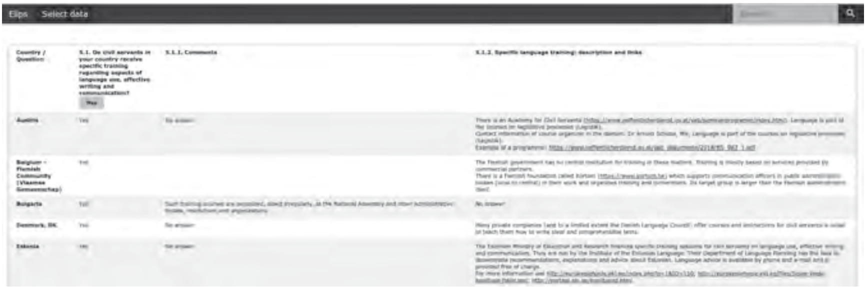
5.1 Methodology

The data collection for ELIPS is based on an online survey conducted in 2018-2019 consisting of 7 main topics and covering 69 different questions. Some are simple yes/no questions while others offer multiple options. As many questions as possible were designed to elicit quantifiable answers which allow for a comparative overview. The comment fields, on the other hand, provide detailed information where nuances and modifications come across. The respondents were invited to provide examples and links which are preserved in the data on the website. Therefore, the comments should always be consulted before drawing conclusions.

5.2 Visualisation

The answers to the questionnaire are displayed on interactive web pages. All questions and answers for all countries can be selected and displayed in a flexible manner. On the ELIPS website (https://elips.efnil.nytud.hu/browse) it is possible to view the answers to all questions for a specific country, to compare the answers to a given question across countries and to combine questions and comments in order to get a more detailed picture.

Comments are given in English. Quotes are given in the original language and in English translation. Active links to current legislation etc. are provided in most cases as shown in Figure 1. Translations of the original quotes are either authorised translations or translations provided by the respondent. This is indicated accordingly.



Country / Question	5.1. Do civil servants in your country receive specific training regarding aspects of language use, effective writing and communication?	5.1.1. Comments	5.1.2. Specific language training: description and links
Austria	Yes	No answer	There is an Academy for Civil Servants ([illegible]). Language is part of the courses on legislative processes (Legistik). Contact information of course organizer in the domain: [illegible]; Language is part of the courses on legislative processes (Legistik). Example of a programme: [illegible]
Belgium – Flemish Community (Vlaamse Gemeenschap)	Yes		The Flemish government has no central institution for training in these matters. Training is mostly based on services provided by commercial partners. There is a Flemish foundation called Kortom (https://www.kortom.be) which supports communication officers in public administration bodies (local to central) in their work and organises training and conventions. Its target group is larger than the Flemish administration itself.
Bulgaria	Yes	Such training courses are organized, albeit irregularly, at the National Assembly and other administrative bodies, institutions and organizations	No answer
Denmark, DK	Yes	No answer	Many private companies (and to a limited extent the Danish Language Council) offer courses and instructions for civil servants in order to teach them how to write clear and comprehensible texts.
Estonia	Yes	No answer	The Estonian Ministry of Education and Research finances specific training sessions for civil servants on language use, effective writing and communication. They are run by the Institute of the Estonian Language. Their Department of Language Planning has the task to disseminate recommendations, explanations and advice about Estonian. Language advice is available by phone and e-mail and is provided free of charge. For more information see [illegible]

Fig. 1: Screenshot of the ELIPS website

For yes/no questions and questions containing quantities, ELIPS offers map views which give a good overview of the results for the participating countries.

The website and its search functions were designed by Ivan Mittelholcz and Ferenczi Zsanett from the Research Institute for Linguistics at the Hungarian Academy of Sciences in cooperation with Sabine Kirchmeier.

6. Results

The following sections present ELIPS topic by topic and summarise the results.

6.1 Plain language policies and actions

The first section of the questionnaire addresses the existence of – and interest in – official plain language policies and the institutions that have been established to implement these policies. It describes explicit policies and measures taken and contains links to language materials, instructions, services and tools available for public administrations. It also touches on how plain-language communication is evaluated and promoted, mapping the degree of international cooperation between official institutions in this field.

6.1.1 Public interest in and institutions for plain language

Clearly, there is public interest in government and public administration using plain language for most languages in the participating countries. Only 7% of the respondents stated that there is no interest and 4% did not know, meaning that 89% said that the use of plain language by government and public administration is indeed a subject of public interest.

Answer	%	Participants
Yes	89%	25
No	7%	2
Unknown	4%	1
Total	100%	28

Table 1: Is the use of plain language by government and public administration a subject of interest in your country?

Consequently, in most countries, there are institutions responsible for maintaining plain language policies and providing plain language services, either the institution of the respondent (29%) or another institution (43%); 14% stated that there are no official institutions while 14% did not know or did not answer the question.

Answer	%	Participants
Yes – my own institution	29%	8
Yes – another institution	43%	12
No	14%	4
Unknown	7%	2
No answer	7%	2
Total	100%	28

Table 2: Is there an institution or body in your country that is responsible for plain language policies for public authorities and/or provides plain-language services for public authorities and/or coordinates the actions of other bodies?

It emerges from the comments that in some countries, like Finland, the subject is well established and plain language policies have existed for about 50 years, whereas in other countries, such as Estonia, the work is just starting. It is evident that at present plain language is not a core activity of EFNIL member institutions and only few of them collaborate with the institutions responsible for that. Fewer than 1/3 of the institutions (8 out of 28) are directly involved in plain language policies, with an additional 3 institutions stating that they collaborate with the plain language institutions. Yet they have knowledge of those institutions' work: 11 respondents named the other institution.

The addresses and links to the institutions responsible for plain language policies in each country can be seen on the ELIPS website (Sections 1.2 and 1.2.2).

6.1.2 Explicit policies and measures for plain language

Recommendations by central governments for government agencies and public administration in general to use plain language were reported by 61% of the participants, with 43% having legal provisions and regulations. More than half of the respondents reported that recommendations exist made by public bodies for their own use. Only one respondent (Lithuania) replied that there are no policies or measures whatsoever for plain language in the country, while 5 respondents did not know or did not answer the question (Austria, Bulgaria, Malta, Portugal and Slovenia).

The most far-reaching provisions can be found in Slovakia and Wales where provisions not only rule that citizens have the right to comprehensible communication by public authorities but also give them the right to refuse unclear information.

Detailed descriptions and links to measures and instructions can be found in Section 1.5 on the ELIPS website.

Country	1.4.1. Recommendations by the central government for government agencies and public administration in general	1.4.2. Legislation by the central government for government agencies and public administration in general	1.4.3. Legal provisions or regulations which rule that citizens have the right to comprehensible communication by public authorities	1.4.4. Legal provisions or regulations that give citizens the right to receive comprehensible communication by public authorities and to refuse unclear information	1.4.5. Recommendations made by separate public administration bodies for their own use
Austria	No Answer	No Answer	No Answer	No Answer	No Answer
Belgium (Flemish Community)	Yes	Yes	No	No	No
Bulgaria	Unknown	Unknown	Unknown	Unknown	Unknown
Denmark	Yes	Yes	No	No	Yes
Estonia	No	No	No	No	Yes
Finland (Swedish)	Yes	Yes	No	No	Yes
Finland (Finnish)	Yes	Yes	No	No	Yes
Germany	No	Yes	No	No	No
Grand Duchy of Luxembourg	Unknown	Yes	Yes	Unknown	Unknown
Greece	Yes	No	No	No	Yes
Hungary	No	No	Yes	No	Yes
Iceland	Yes	Yes	No	No	No
Ireland (excl. Northern Ireland)	Yes	No	No	No	No
Italy	Yes	No	No	No	Yes
Latvia	Yes	Yes	Yes	No	Yes
Lithuania	No	No	No	No	No

Country	1.4.1. Recommendations by the central government for government agencies and public administration in general	1.4.2. Legislation by the central government for government agencies and public administration in general	1.4.3. Legal provisions or regulations which rule that citizens have the right to comprehensible communication by public authorities	1.4.4. Legal provisions or regulations that give citizens the right to receive comprehensible communication by public authorities and to refuse unclear information	1.4.5. Recommendations made by separate public administration bodies for their own use
Malta	No Answer	No Answer	No Answer	No Answer	No Answer
Netherlands	Yes	No	No	No	Yes
Norway	Yes	No	No	No	Yes
Portugal	No Answer	No Answer	No Answer	No Answer	No Answer
Slovak Republic	Yes	Yes	Yes	Yes	Yes
Slovenia	No Answer	No Answer	No Answer	No Answer	No Answer
Sweden	Yes	Yes	No	No	Yes
Switzerland	Yes	Yes	Yes	No	Yes
UK (England)	Yes	No	No	No	Yes
UK (Wales)	Yes	Yes	Yes	Yes	Yes
UK (Northern Ireland)	Yes	No	No	No	Yes
UK (Scotland)	No	No	No	No	Yes

Table 3: Explicit policies or policy measures or instructions addressing the use of plain language within public administration

6.1.3 Plain language materials, services and tools

Regarding methods to help public administrations comply with the principles of plain language, the publication of guidelines seems to be the most widespread. Three quarters (21 out of 28 respondents) reported that such measures are used. Web services also seem rather popular (used by 68%) while 36% mentioned the use of templates and 39% the use of digital tools such as style checkers or complexity-of-text predictors. Public administrations in Denmark, Finland (those working in Finnish), Greece, Norway and Sweden seem to have the whole palette of possibilities available.

	1.6. Materials, instructions, services and tools			
Country	Web service(s)	Guidelines (online, pdf or printed)	Models or templates	Tools
Austria	No Answer	No Answer	No Answer	No Answer
Belgium (Flemish Community)	Yes	Yes	Yes	Yes
Bulgaria	Yes	No Answer	No Answer	No Answer

	1.6. Materials, instructions, services and tools			
Country	Web service(s)	Guidelines (online, pdf or printed)	Models or templates	Tools
Denmark	Yes	Yes	Yes	Yes
Estonia	No Answer	Yes	No Answer	No Answer
Finland (Swedish)	Yes	Yes	Yes	No Answer
Finland (Finnish)	Yes	Yes	Yes	Yes
Germany	Yes	Yes	No Answer	Yes
Grand Duchy of Luxembourg	Yes	Yes	Yes	No Answer
Greece	Yes	Yes	Yes	Yes
Hungary	No Answer	Yes	No Answer	No Answer
Iceland	Yes	Yes	No Answer	No Answer
Ireland (excl. Northern Ireland)	Yes	Yes	No Answer	Yes
Italy	Yes	Yes	No	No
Latvia	Yes	No Answer	No Answer	No Answer
Lithuania	Yes	Yes	No Answer	Yes
Malta	No Answer	No Answer	No Answer	No Answer
Netherlands	Yes	Yes	No Answer	Yes
Norway	Yes	Yes	Yes	Yes
Portugal	No Answer	No Answer	No Answer	No Answer
Slovak Republic	Yes	Yes	No Answer	No Answer
Slovenia	Yes	No Answer	No Answer	Yes
Sweden	Yes	Yes	Yes	Yes
Switzerland	No Answer	Yes	No Answer	No Answer
UK (England)	Yes	Yes	Yes	No Answer
UK (Wales)	No Answer	Yes	Yes	No Answer
UK (Northern Ireland)	Unknown	Unknown	Unknown	Unknown
UK (Scotland)	No Answer	Yes	No Answer	No Answer

Table 4: Which materials, instructions, services and tools are available in your country in order to help public administration comply with the principles of plain language?

Descriptions of and links to materials, instructions, services and tools are available in Section 1.7 on the ELIPS website.

6.1.4 Endeavours to measure the effect of plain language policies

One third of the respondents reported that there are projects that aim to measure the effect of plain language policies either in terms of increased quality and user satisfaction or in terms of efficiency. Authorities in Norway have developed an online toolbox with methods for user involvement and measuring results and in the Netherlands a proposal has been submitted for a project that aims to monitor plain language results. Only Finland referred to documented studies, with other countries mainly referring to projects in progress (cf. Section 1.8 on the ELIPS website).

6.1.5 Promotion of plain language policies and awareness

Just over half, or 54%, of the respondents reported that there are initiatives to promote plain language policies in their country. The strategies range from launching a plain language prize to competitions and campaigns. Awards are given for different achievements, for instance, the clearest text, the best author or the best promoter of plain language. In Wales, it is possible to obtain a quality seal if certain conditions are met.

Detailed descriptions and links to various initiatives can be found in Section 1.8 on the ELIPS website.

6.1.6 International cooperation

Estonia, Finland, Norway, the Slovak Republic and Sweden said that they are members of one or both of the two main international organisations for plain language, *PLAIN* and *Clarity*. Six other respondents reported their involvement in other organisations or conferences. About half of the respondents are not involved in any kind of international cooperation.

Country	1.10.1. Member of PLAIN	1.10.2. Member of Clarity	1.10.3. Member of other organisations	1.10.4. Involvement in international conferences	1.10.5. Involvement in other types of international cooperation
Austria	No Answer	No Answer	No Answer	No Answer	No Answer
Belgium (Flemish Community)	No	No	Yes	No	No
Bulgaria	No	No	No	No	No
Denmark	No	No	No	Yes	No
Estonia	No	Yes	No	No	No
Finland (Swedish)	Yes	Yes	Yes	Yes	No
Finland (Finnish)	Yes	Yes	Yes	Yes	No
Germany	No	No	No	No	No
Grand Duchy of Luxembourg	No	No	No	No	No
Greece	No	No	No	No	No
Hungary	No	No	No	No	No
Iceland	No	No	Yes	Yes	No

Country	1.10.1. Member of PLAIN	1.10.2. Member of Clarity	1.10.3. Member of other organisations	1.10.4. Involvement in international conferences	1.10.5. Involvement in other types of international cooperation
Ireland (excl. Northern Ireland)	No	No	No	No	No
Italy	No	No	Yes	No	No
Latvia	No	No	No	No	No
Lithuania	No	No	Yes	Yes	Yes
Malta	No	No	No	No	No
Netherlands	No	No	Yes	No	No
Norway	Yes	Yes	Yes	Yes	No
Portugal	No	No	No	No	No
Slovak Republic	Yes	Yes	No	No	No
Slovenia	No	No	No	No	No
Sweden	Yes	Yes	Yes	Yes	No
Switzerland	No	No	No	No	No
UK (England)	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Wales)	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Northern Ireland)	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Scotland)	Unknown	Unknown	Unknown	Unknown	Unknown

Table 5: Is your institution involved in international cooperation concerning plain language?

Descriptions and links to various plain language organisations, networks and conferences can be found in Section 1.11 on the ELIPS website.

6.2 Easy-to-read language policies and actions

The basic difference between easy-to-read language and plain language is the target audience. Whereas easy-to-read language texts specifically address persons with reading or comprehension barriers, plain language texts address the public reader in general.

More than half of the respondents (53%) confirmed the existence of legislation or recommendations by central government agencies and public administration in general. Almost one third (29%) reported on the existence of recommendations made by separate public administration bodies for their own use while 57% seemed to have nothing of the kind.

Country	2.1.1. Legislation or recommendations by public administrations in general	2.1.2. Recommendations made by separate public administration bodies for their own use
Austria	Yes	No
Belgium (Flemish Community)	No	Yes
Bulgaria	Yes	Yes
Denmark	No	No
Estonia	No	No
Finland (Swedish)	Yes	Yes
Finland (Finnish)	Yes	Yes
Germany	Yes	No
Grand Duchy of Luxembourg	Unknown	Unknown
Greece	Yes	No
Hungary	Yes	No
Iceland	Yes	Yes
Ireland (excl. Northern Ireland)	No	Yes
Italy	No	No
Latvia	Yes	No
Lithuania	Unknown	Unknown
Malta	No	No
Netherlands	No	No
Norway	Yes	Yes
Portugal	No	No
Slovak Republic	Yes	No
Slovenia	No	No
Sweden	Yes	No
Switzerland	Yes	No
UK (England)	Yes	Yes
UK (Wales)	Unknown	Unknown
UK (Northern Ireland)	No	Yes
UK (Scotland)	Yes	No

Table 6: Are there explicit policies or policy measures or instructions addressing the use of easy-to-read language in some cases for some target groups?

The respondents provided a number of references to local or global guidelines, such as to the recommendations of the *World Wide Web Consortium*.[14] The references can all be found in Sections 2.1.3 to 2.2.2 on the ELIPS website.

[14] See website: https://www.w3.org/TR/WCAG20/.

In some countries, there are separate guidelines for easy-to-read language, whereas in others, the guidelines are part of the guidelines for plain language. A few countries are still working out policies in this field.

Just over one fifth (22%) of the respondents reported that there is an institution or body responsible for the use of easy-to-read languages by public institutions. In most cases (18%), it is not the respondents' own institution but some other body or institution. It is noteworthy that the largest group of respondents (32%) did not answer this question.

No	29%
No answer	32%
Unknown	18%
Yes, another institution than the respondent's	18%
Yes, the respondent's institution	4%

Table 7: Is there an institution or body that is responsible for the use of easy-to-read language by public authorities and/or provides easy-to-read language services for public authorities and/or coordinates the actions of other bodies?

Detailed information and links about institutions dedicated to working with easy-to-read language can be found in Section 2.4 on the ELIPS website.

6.3 Terminology policies and actions

6.3.1 Public interest in terminology

The interest in terminology seems to be quite strong in the participating countries and is well known to the responding institutions: 86% reported that the use of terminology within government and public administration is a subject of public interest.

No	11%
Unknown	4%
Yes	85%

Table 8: Is (the use of) terminology within government and public administration a subject of public interest in your country?

In all, 29% of the participants stated that the responsibility for terminology development and/or terminology policies lies within the respondent's own institution and 39% reported that there are other institutions that deal with terminology. In these cases, most of the respondents' institutions collaborate directly or in some other way. Around one fifth (21%) reported that there are no institutions responsible for terminology management.

Descriptions of the collaboration and links to other terminology institutions can be found in Section 3.2 on the ELIPS website.

6.3.2 Terminology management tools

The respondents were also asked to provide information about which methods are used to help public institutions with the acceptance, use and description of terminology. Here, terminology databases and terminology extraction tools turned out to be the most widely used, with 68% (19 out of 28) of the respondents indicating that terminology databases and extraction tools are used. In addition, 57% (16 out of 28) of the respondents stated that official guidelines, legal acts or regulations are in use and 46% (13 out of 28) reported that web services are used.

Country	3.3.1. Web service(s)	3.3.2. Official guidelines, legal acts or regulations	3.3.3. Tools
Austria	No	No	No
Belgium (Flemish Community)	Yes	Yes	Yes
Bulgaria	Unknown	Unknown	Unknown
Denmark	No	No	Yes
Estonia	No	No	Yes
Finland (Swedish)	Yes	Yes	Yes
Finland (Finnish)	Yes	Yes	Yes
Germany	No	No	No
Grand Duchy of Luxembourg	No	No	No
Greece	Yes	Yes	Yes
Hungary	No	No	No
Iceland	Yes	Yes	Yes
Ireland (excl. Northern Ireland)	Yes	Yes	Yes
Italy	Unknown	Unknown	Unknown
Latvia	Yes	Yes	Yes
Lithuania	Yes	Yes	Yes
Malta	No	Yes	No
Netherlands	Yes	No	Yes
Norway	Yes	Yes	Yes
Portugal	No	No	No
Slovak Republic	Yes	Yes	Yes
Slovenia	Yes	No	Yes
Sweden	No	No	Yes
Switzerland	No	Yes	Yes
UK (England)	No	Yes	No

Country	3.3.1. Web service(s)	3.3.2. Official guidelines, legal acts or regulations	3.3.3. Tools
UK (Wales)	Yes	Yes	Yes
UK (Northern Ireland)	No	Yes	Yes
UK (Scotland)	No	Yes	Yes

Table 9: Which of the following specific materials, instructions, services and tools are available in your country in order to help public administration with the acceptance, use and description of terminology?

Detailed information and links to guidelines, tools and web services can be found in Section 3.4 on the ELIPS website.

6.3.3 International cooperation about terminology

Although there seems to be strong interest in terminology in almost all countries, international cooperation on terminology is not equally widespread. Furthermore, those countries that do collaborate internationally do not use the same conferences or networks so the picture is rather diverse. Some countries are associated with the *European Association for Terminology* (EAFT-AET), a few with *TermNET* and only one, the Slovak Republic, reported that it makes use of the *Conference of Translation Services of European States* (COTSOES).

In all 39% of the respondents stated that their institutions are members of other conferences or networks. For instance, many of the Nordic countries are organised in *Nordterm* and others are associated with the *EuroTermBank* project that runs under the EU's *Connecting Europe Facility* (CEF).

Country	3.5.1. European Association for Terminology (EAFT-AET)	3.5.2. TermNet	3.5.3. Conference of Translation Services of European States	3.5.4. Other international organisations or networks	3.5.5. International conferences and symposia	3.5.6. Other forms of collaboration
Austria	No	Yes	No	No	No	No
Belgium (Flemish Community)	Yes	Yes	No	No	No	No
Bulgaria	No	No	No	Yes	No	No
Denmark	No	No	No	No	Yes	No
Estonia	Yes	No	No	No	No	No
Finland (Swedish)	No	No	No	No	Yes	No
Finland (Finnish)	No	No	No	No	Yes	No
Germany	No	No	No	No	No	No
Grand Duchy of Luxembourg	No	No	No	No	No	No
Greece	No	No	No	No	No	No
Hungary	No	No	No	No	No	No

Country	3.5.1. European Association for Terminology (EAFT-AET)	3.5.2. TermNet	3.5.3. Conference of Translation Services of European States	3.5.4. Other international organisations or networks	3.5.5. International conferences and symposia	3.5.6. Other forms of collaboration
Iceland	Yes	Yes	No	Yes	Yes	No
Ireland (excl. Northern Ireland)	Yes	No	No	Yes	Yes	No
Italy	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Latvia	No	No	No	Yes	No	No
Lithuania	Yes	Yes	No	Yes	Yes	Yes
Malta	No	No	No	No	No	No
Netherlands	No	No	No	No	No	Yes
Norway	Yes	No	No	Yes	Yes	No
Portugal	No	No	No	No	No	No
Slovak Republic	Yes	Yes	Yes	Yes	Yes	No
Slovenia	No	No	No	Yes	No	No
Sweden	No	No	No	No	No	No
Switzerland	No	No	No	No	No	No
UK (England)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Wales)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Northern Ireland)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Scotland)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown

Table 10: Is your institution involved in international cooperation concerning terminology?

Descriptions and links to other conferences and networks that are used can be found in Section 3.6 on the ELIPS website.

6.4 Policies and actions on the use of other languages as well as gender, cultural and sexual diversity

Just over two thirds of the respondents (68%) indicated that there are language-specific instructions or guidelines for communication by public authorities for using languages other than official languages, for instance minority languages, foreign languages or sign language, in certain cases and for certain target groups. Rulings for minority languages such as Sámi and sign languages are very prominent in this group.

A slightly smaller group (64%) stated that there are official guidelines on the use of gender-neutral language and other gender aspects such as the masculine and feminine forms for the names of functions and titles.

Language-specific instructions or guidelines on cultural diversity and/or sexual preferences seem to be less widespread (29%). In Sweden, such research projects have only been initiated recently. In the UK, these issues are covered by the guidelines for gender equality and Italy has guidelines for non-sexist language as well.

Other issues include disabilities, mental health, religion, nationality and age. Nearly one third (29%) of the respondents indicated that there are guidelines on such other issues as well.

Country	4.1.1. Guidelines on the use of other languages (minority languages, foreign languages, sign language)	4.1.2. Guidelines on the use of gender-neutral language and other gender aspects	4.1.3. Guidelines on culturasl diversity and/or sexual preferences	4.1.4. Guidelines on other issues
Austria	Yes	Yes	Yes	No
Belgium (Flemish Community)	No	Yes	No	No
Bulgaria	Unknown	Unknown	Unknown	Unknown
Denmark	No	No	No	Yes
Estonia	No	No	No	No
Finland (Swedish)	Yes	Yes	Yes	No
Finland (Finnish)	Yes	Yes	Yes	No
Germany	Yes	Yes	No	No
Grand Duchy of Luxembourg	No	No	No	No
Greece	Yes	Yes	No	No
Hungary	Yes	Yes	Yes	Yes
Iceland	Yes	No	No	No
Ireland (excl. Northern Ireland)	Unknown	Unknown	Unknown	Unknown
Italy	Yes	Yes	Unknown	Unknown
Latvia	No	No	No	No
Lithuania	Yes	Unknown	Unknown	Yes
Malta	Unknown	Unknown	Unknown	Unknown
Netherlands	Yes	Yes	No	No
Norway	Yes	Yes	No	No
Portugal	Yes	No	No	No
Slovak Republic	Yes	Yes	Yes	Yes
Slovenia	No	Yes	No	No
Sweden	Yes	Yes	Yes	No
Switzerland	Yes	Yes	No	No
UK (England)	Yes	Yes	Yes	Yes
UK (Wales)	Yes	Yes	Yes	No
UK (Northern Ireland)	Yes	Yes	No	No
UK (Scotland)	Yes	Yes	No	No

Table 11: Are there other language-specific instructions or guidelines for communication by public authorities in your country?

Only 11% of the respondents indicated that plain language principles also apply to guidelines and instructions for other languages and special groups. However, in many countries there may be the same attitude as in Switzerland, where the response is as follows: “In principle, all publicly available information issued by federal authorities is subject to the same principles. There is no explicit mention in the relevant laws, by-laws or guidelines that some languages would be exempt from this principle”.

Detailed descriptions and links to national guidelines and instructions can be found in Section 4.2 on the ELIPS website.

6.5 Training

Just over two thirds (68%) of the respondents replied that civil servants receive specific training regarding aspects of language use, effective writing and communication. Of course, quite a number of linguistic aspects can be addressed, as can be seen in Figure 2.

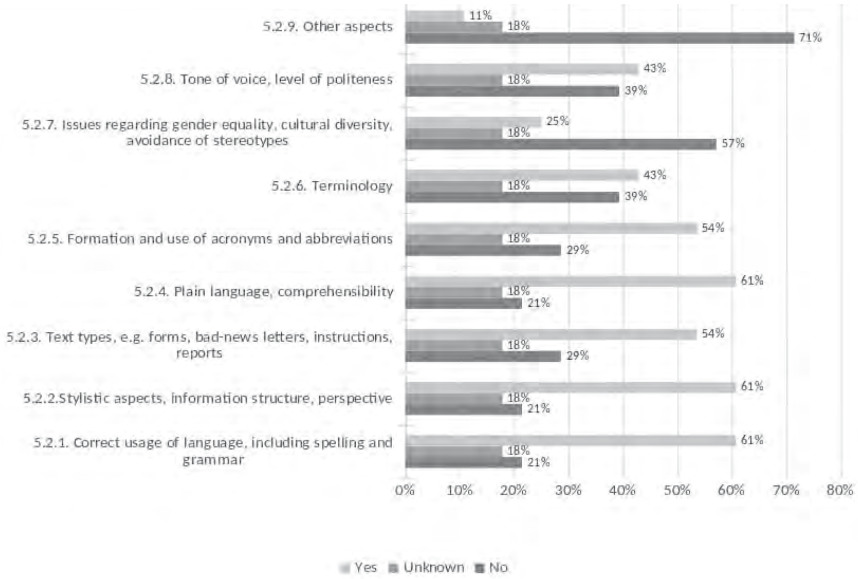


Fig. 2: What aspects are addressed in training? (Summary)

In fact, it seems that most aspects are addressed in training, although terminology and tone of voice seem to receive a little less attention. These topics were only mentioned by 43% of the respondents, while over half of them reported training for most other domains. The least prominent domains regard gender equality, cultural diversity and avoidance of stereotypes, being mentioned by only 25% of the respondents.

Country	5.2.1. Correct usage of language, including spelling and grammar	5.2.2. Stylistic aspects, information structure, perspective	5.2.3. Text types, e.g. forms, bad-news letters, instructions, reports	5.2.4. Plain language, comprehensibility	5.2.5. Formation and use of acronyms and abbreviations	3.5.6.Terminology	3.5.7. Issues regarding gender equality, cultural diversity, avoidance of stereotypes	3.5.8. Tone of voice, levels of politeness	3.5.9. Other aspects
Austria	Yes	No	No	Yes	Yes	Yes	Yes	No	No
Belgium (Flemish Community)	Yes	Yes	Yes	Yes	Yes	No	No	Yes	No
Bulgaria	Yes	Yes	No	No	No	No	No	No	No
Denmark	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Estonia	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Finland (Swedish)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Finland (Finnish)	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Germany	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Grand Duchy of Luxembourg	Yes	Yes	Yes	Yes	Yes	No	No	No	No
Greece	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Hungary	Yes	Yes	No	Yes	Yes	No	No	No	No
Iceland	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Ireland (excl. Northern Ireland)	No	No	No	No	No	No	No	No	No
Italy	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
Latvia	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Lithuania	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Malta	Yes	Yes	Yes	Yes	No	No	No	Yes	No
Netherlands	No	No	No	No	No	No	No	No	Yes
Norway	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes	No
Portugal	No	No	No	No	No	No	No	No	No
Slovak Republic	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No
Slovenia	Yes	Yes	Yes	Yes	No	No	No	No	No
Sweden	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Switzerland	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
UK (England)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Wales)	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown
UK (Northern Ireland)	No	No	No	No	No	No	No	No	No
UK (Scotland)	No	No	No	No	No	No	No	No	No

Table 12: What aspects are addressed in training?

Among the other topics addressed in the linguistic training of civil servants are the formation of plain official proper names and general communication skills.

Descriptions and links to training principles and training facilities can be found in Section 5.2 on the ELIPS website.

6.6 Collaboration between member states and the EU

The question of international collaboration has already been addressed several times in the previous sections. In this section, however, we specifically focus on collaboration between member states and the EU. Half of the respondents stated that there is some kind of formal collaboration platform that links the language services of the EU with the official institutions for language. The rest answered negatively or simply did not know.

Country	6.1. Is there a platform for collaboration and coordination between the language services of the EU and the national institutions regarding your national language(s)?	6.3. Is your institution involved in the collaboration platform?
Belgium (Flemish Community)	Yes	No
Denmark	Yes	No
Estonia	Yes	Yes
Finland (Finnish)	Yes	Yes
Greece	Yes	Yes
Hungary	Yes	Yes
Ireland (excl. Northern Ireland)	Yes	No
Italy	Yes	Yes
Netherlands	Yes	Yes
Portugal	Yes	Yes
Slovak Republic	Yes	No
Switzerland	Yes	No
UK (Northern Ireland)	Yes	No
Lithuania	Yes	Not relevant

Table 13: Collaboration with language services of the EU

For those countries that do have formalised collaboration, the main issues addressed were translation tools, terminology databases and tools. Collaboration on plain language was reported in 5 cases and exchanges about gender equality and cultural diversity were only reported for Italy.

	6.2. What aspects do the platforms address?					
Country	6.2.1. Translation tools (dictionaries, corpora, translation memories etc.)	6.2.2. Terminology bases and tools, e.g. for terminology extraction	6.2.3. Plain language and comprehensibility	6.2.4. Gender equality and cultural diversity	6.2.5. Style guides, templates, models	6.2.6. Organisation of meetings, conferences and training sessions
Belgium (Flemish Community)	No	Yes	No	No	No	No
Denmark	Yes	Yes	No	No	No	No
Estonia	Yes	Yes	Yes	No	Yes	Yes
Finland (Finnish)	No	Yes	No	No	No	No
Greece	No	Yes	No	No	No	No
Hungary	No	No	Yes	No	Yes	Yes
Ireland (excl. Northern Ireland)	Yes	No	No	No	No	No
Italy	Yes	Yes	Yes	Yes	No	No
Netherlands	No	Yes	No	No	No	No
Portugal	Yes	No	No	No	No	No
Slovak Republic	Yes	Yes	Yes	No	No	Yes
Switzerland	Yes	Yes	No	No	Yes	Yes
UK (Northern Ireland)	Yes	No	No	No	No	No
Lithuania	Yes	Yes	Yes	Not relevant	Not relevant	Yes

Table 14: Domains of collaboration with language services of the EU

7. Conclusions and recommendations

7.1 Conclusions

The analysis of the answers as described in the paragraphs above show that most of the participating countries do have policies related to the use and quality of their (national) languages as instruments for government, legislation and public administration. Many of these policies also cover the various aspects which were the focus of our ELIPS survey.

However, there seem to be large differences in the attention paid to the various subdomains. Terminology and plain language seem to receive the most widespread attention. Fields such as easy-to-read language as well as social, cultural and gender diversity are less well established and/or seem to be more recent, probably as a result of an increasing sensitivity towards these aspects over the last few years as they are considered constituents of inclusive communication. Moreover, even well-established fields show important impact differences between the countries which participated in the survey. In Finland and Sweden, for instance, plain language policies have existed for about 50 years or so, while many other countries like Estonia and the Netherlands have only started working on them recently.

The answers to the survey also show that, as a rule, policies are developed on a national scale without too much awareness of what other languages and countries do, to say nothing of active interchange or cooperation. Most countries are not involved in international organisations and networks such as *PLAIN* (Plain Language Association International) and *Clarity* for plain language or *EAFT* (European Association for Terminology) or *COTSOES* (Conference of Translation Services of European States) as far as international platforms for terminology are concerned.

If we look at the various subdomains within the field of the institutional use of languages, we also see that these national policies are fragmented. There is no coherence and almost no exchange or collaboration between the various subdomains and bodies responsible for it, e.g. between plain language and easy-to-read actors, or between official terminology bodies and actors in the field of diversity.

Our survey also brings us to a third observation: the discontinuity between the level of the nation state and the institutions of the European Union. Typically, EU institutions are not involved or consulted in the definition and evaluation of language-specific policies, even though the quality of European regulations has a direct influence on public communication on a national level because member states have to integrate European rulings into their national legislation.

This leads us to the conclusion that more coherence and convergence between the various domains, a better sharing of experiences and practices between the various nation states in Europe and more continuity and interaction between national and European policy levels could be beneficial for the overall quality and effectiveness of language use within the domains of government, legislation and public administration.

Last but not least, the survey gives us a good idea of the involvement of the member institutions of EFNIL in these official language policies. Many EFNIL members have a direct commitment and involvement in the policies addressed by this survey, either as primary actors responsible for some or even all of these fields or as collaborating parties with the institutions that are directly in charge, while some members have no involvement whatsoever. The degree of involvement differs from country to country and from subdomain to subdomain. It seems strongest for terminology, followed by plain language.

This leads us to the conclusion that there are various opportunities for EFNIL to be instrumental in strengthening these policies and contributing to more coherence and comparability within Europe as a whole, e.g. by encouraging members from countries with weaker or absent policies to help their country close the gap and, in doing so, build on the experiences of colleagues in countries with strong traditions and active policies or by encouraging its members to act as intermediaries between subdomains and between national and European levels in order to stimulate cooperation and strengthen overall coherence. This leads us to a number of recommendations to EFNIL and EFNIL member institutions alike which are included in the next few paragraphs. Although these recommendations focus on EFNIL and EFNIL member institutions, we sincerely hope that both the survey and our conclusions and recommendations will prove to be useful and inspiring to all other users, for instance to academic experts when identifying topics for research or to governments and policy bodies when comparing their national situation with other countries and even to identify partners for international cooperation.

7.2 Recommendations

7.2.1 Recommendations for member institutions about national activities

The ELIPS group recommends that EFNIL member institutions consider the following actions:

1) The member institutions could involve themselves more in national plain language and easy language activities to strengthen their position as national expert institutions, for example:
   - If there is a national body responsible for that, member institutions could organise joint conferences with that body about themes that are common to both or connect to the core activities of each (e.g. the translation of communications by public authorities into national minority languages and the quality of those texts). They could also carry out joint projects or lobby together for the creation of national policies or influence their content.
   - The member institutions could convene national actors from several different domains (e.g. plain language, easy language and terminology actors as well as actors promoting inclusive policies) and bring them together at conferences or meetings to examine the possibilities of promoting their domains together or forming national policies for them, e.g. language as a part of accessibility policies.
   - If no body exists for any given domain, the member institutions could bring together individual actors in one or several such domains (plain language, easy language, gender neutral language, inclusive language) and offer a platform to exchange best practices and find common goals of action.
2) The member institutions could participate more often in international cooperation on plain language, easy language, terminology and other domains, i.e. joining international organisations and participating in international conferences in the relevant field to exchange experiences and best practices and to benefit from them.

3) The member institutions could get involved in developing and localising the international ISO standard for plain language in a national standard via the national standardisation organisations to lend their expertise and gain networks for their own tasks.

7.2.2 Recommendations for EFNIL as an organisation, influencing outwards and continuation of the project

The ELIPS group recommends that EFNIL considers the following suggestions:

1) EFNIL could organise conferences and meetings for its member institutions and outside experts about plain language, easy language and other domains of the survey in order to exchange experiences and best practices and to provide opportunities for partnerships and networking for those involved or interested in the same fields of activities. Strengthening especially those domains that receive less attention at present (especially gender neutrality and inclusive language) would enhance the overall quality and suitability of the language use by public authorities in member countries.
   – One theme for conferences could be the impact and effectiveness of plain language, easy-to-read and diversity policies since in many countries there is a need to demonstrate the return on investments in these. The conference could present findings on the effects of completed projects, both in material terms (reduction in costs, e.g. as a result of fewer complaints, legal actions etc.) and in immaterial terms (increased trust in institutions) and discuss their reliability.
   – Another theme could be the possible benefits of integrating national language resources (terminology collections, translation memories etc.) in a multilingual language infrastructure. Many EFNIL member institutions seem to be directly involved in policies and corpus planning regarding (legislative and administrative) terminology for their language. Cooperation with EU terminology experts and the IATE database would be beneficial to all parties.
2) EFNIL could commission or initiate a comparative review of tools for plain language and easy language which are already in use. International collaboration on sharing the same or comparable technological and linguistic bases for these tools can lead to a considerable gain in quality. It could also help develop comparable tools for those languages where such tools are not yet available. EFNIL could also contact universities or research institutes in member countries with research in these fields to sound out their interest in a research project which could apply for EU project funding.
3) EFNIL could explore with the European Commission (and perhaps also with the Secretariat of the Parliament and/or the Council of the EU) the possibility

of convening relevant national actors in different domains examined in the survey (e.g. competent bodies or other experts) to discuss whether common recommendations can be formulated for establishing national policies to promote plain language, easy language and other forms of inclusive use of language (not texts but procedures, tools, institutions).

References

Araújo, L./Dins da Costa, P. (2013): *The European Survey on Language Competences*. Commissioned by CRELL Centre for Research on Education and Lifelong Learning. JRC 82366.

Cheek, A. (2010): Defining plain language. In: *Clarity Journal* 64, 5-15. https://www.clarity-international.org/wp-content/uploads/2020/07/Clarity-no-64-bookmarked1.pdf.

CWU – the Communications Union (2013): *Use of English in the workplace – a guide*. http://www.cwu.org/media/6683/cwu__1346163255_use_of_language_in_the_workpla.pdf.

Ehrenberg-Sundin, B./Sundin, M. (2015): Krångelspråk blir klarspråk – från 1970-tal till 2010-tal. (= Språkrådets skrifter 18). Stockholm: Norstedts.

Hansson, K. (2020): “*Det finns ett sug efter klarspråk*”. *En studie om bättre stöd till klarspråk i offentlig verksamhet*. (= Rapporter från Språkrådet 16). Stockholm: Institutet för språk och folkminnen. https://isof.diva-portal.org/smash/get/diva2:1440648/FULLTEXT01.pdf.

International Plain Language Federation: *Plain Language Definitions*. https://www.iplfederation.org/plain-language.

Kimble, J. (2016): A curious criticism of plain language. In: *Legal Communication & Rhetoric, JALWD* 13, 181-191. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2881592.

Kirchmeier-Andersen, S./Robustelli, C./Spetz, J./Stickel, G./Teigland, N. (2010): European Language Monitor (ELM). In: Stickel, G. (ed.): *National, regional and minority languages in Europe. Contributions to the Annual Conference 2009 of EFNIL in Dublin*. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 81). Frankfurt a. M.: Peter Lang, 181-190.

Kokoroskos, D. (2012): European Language Monitor (ELM): interim report. In: Stickel, G. (ed.): *National, regional and minority languages in Europe. Contributions to the Annual Conference 2009 of EFNIL in Dublin*. (= Duisburger Arbeiten zur Sprach- und Kulturwissenschaft 81). Frankfurt a. M.: Peter Lang, 191-217.

Lindholm, C./Vanhatalo, U. (eds.) (2021): *Handbook of easy languages in Europe*. Berlin: Frank & Timme. https://library.oapen.org/bitstream/handle/20.500.12657/52628/Handbook_of_Easy_Languages_in_Europe.pdf?sequence=12&isAllowed=y.

Martinho, M. (2017): *International public sector survey*. Paper presented at 11th PLAIN Conference, Graz, 23 September 2017. http://plainlanguagenetwork.org/wp-content/uploads/2017/11/63.-Miguel-Martinho-International-public-sector-survey.pdf.

Mikhailov, M./Piehl, A. (2018): Observing Eurolects: the case of Finnish. In: Mori, L. (ed.): *Observing Eurolects. Corpus analysis of linguistic variation in EU law*. (= Studies in Corpus Linguistics 86). Amsterdam/Philadelphia. Benjamins.

Mori, L.. (ed.) (2018): *Observing Eurolects. Corpus analysis of linguistic variation in EU law*. (= Studies in Corpus Linguistics 86). Amsterdam/Philadelphia. Benjamins.

Piehl, A. (2008): Finland makes its statutes intelligible: Good intentions and practicalities. In: Wagner, A./Cacciaguidi-Fahy, S. (eds.): Obscurity and clarity in the law. Prospects and challenges. Farnham: Ashgate. https://www.kotus.fi/files/2072/Obscurity_and_Clarity_in_the_Law.pdf.

Piehl, A. (2019): Plain Finnish in the European Union: mission possible? In: The Clarity Journal 80, 45-47. https://www.clarity-international.org/wp-content/uploads/2019/06/Clarity_80.pdf.

Schriver, K. (2017): Plain language in the US Gains Momentum: 1940-2015. In: IEEE Transactions on Professional Communication 60, 4 (December 2017), 343-383.

Somssich, R./Várnai, J./Bérczi, A. (2010): *Study on lawmaking in the EU multilingual environment*. (= Studies on Translation and Multilingualism 1/2010). Bruxelles: European Commission, Directorate-General for Translation. https://www.academia.edu/37495191/European_Commission_Directorate_General_for_Translation_Studies_on_translation_and_multilingualism_Lawmaking_in_the_EU_multilingual_environment_Lawmaking_in_the_EU_multilingual_environment_1_2010.

Tiililä, U. (2018): Virkakielityön periaatteet: työtä kielen parissa ihmisten hyväksi. In: *Kielikello* 4. https://www.kielikello.fi/-/virkakielityon-periaatteet-tyota-kielen-parissa-ihmisten-hyvak-1.

Viertiö, A. (2011): *Hallinnossa kaivataan koulutusta ja laatua viestintään*. In: *Kielikello* 4. https://www.kielikello.fi/-/hallinnossa-kaivataan-koulutusta-ja-laatua-viestintaan.

Appendix

European Federation of National Institutions for Language (EFNIL): Member institutions

For detailed information on EFNIL and its members see www.efnil.org

Member institutions grouped by country

Austria	***Österreichisches Sprachen-Kompetenz-Zentrum***, Graz Austrian Centre for Language Competence ***Austrian Centre for Digital Humanities***, ***Österreichische Akademie der Wissenschaften***, Wien/Vienna Austrian Academy of Sciences
Belgium	***Ministère de la Fédération Wallonie Bruxelles Service de la Langue française***, Bruxelles/Brussels Federation Wallonia-Brussels
Bulgaria	***Българска академия на науките, Институт за български език***, Sofia Bulgarian Academy of Sciences, Institute for the Bulgarian Language
Croatia:	***Institut za hrvatski jezik i jezikoslovlje***, Zagreb Institute of Croatian Language and Linguistics
Czech Republic	***Ústav pro jazyk český Akademie Věd České republiky, v. v. i.***, Praha/Prague Czech Language Institute of the Czech Academy of Sciences
Denmark	***Dansk Sprognævn***, København Danish Language Council
Estonia	***Eesti Keele Instituut***, Tallin Institute of the Estonian Language ***Eesti Keelenõukogu***, Tallin Estonian Language Council
Finland	***Kotimaisten kielten keskus, Institutet för de inhemska språken***, Helsinki/Helsingfors Institute for the Languages of Finland

France	***Délégation Générale à la langue française et aux langues de France***, Paris General Delegation for the French Language and the Languages of France
Georgia	***Tbilisi State University***, Tbilisi State Language Department
Germany	***Leibniz-Institut für Deutsche Sprache***, Mannheim Leibniz-Institute for the German Language ***Deutsche Akademie für Sprache und Dichtung***, Darmstadt German Academy for Language and Literature
Greece	***Κέντρο Ελληνικής Γλώσσας/Kentro Ellinikis Glossas***, Thessaloniki Centre for the Greek Language
Hungary	***Nyelvtudományi Kutatóközpont***, Budapest Hungarian Research Centre for Linguistics
Ireland	***Foras na Gaeilge***, Dublin (the all-island body for the Irish language)
Iceland	***Stofnun Árna Magnússonar í íslenskum fræðum***, Reykjavik The Árni Magnússon Institute of Icelandic Studies
Italy	***Accademia della Crusca***, Firenze/Florence (the central academy for the Italian language) ***CNR – Opera del Vocabolario Italiano***, Firenze/Florence Italian Dictionary Institute
Latvia	***Latviešu valodas institūts***, Riga Latvian Language Institute ***Latviešu valodas aġentūra***, Riga State Language Agency
Lithuania	***Lietuvių Kalbos Institutas***, Vilnius Institute of the Lithuanian Language ***Valstybiné Lietuvių Kalbos Komisija***, Vilnius The State Commission of the Lithuanian Language
Luxembourg	***Institut Grand-Ducal***, Luxembourg Grand Ducal Institute

	Zenter fir d'Lëtzebuerger Sprooch vum Ministère fir Educatioun, Kanner a Jugend, Luxembourg (Center for the Luxembourgish Language of the Ministry of Education, Children and Youth)
Malta	***Il-Kunsill Nazzjonali tal-Ilsien Malti***, Floriana National Council for the Maltese Language
Netherlands	***Instituut voor de Nederlandse Taal***, Leiden Dutch Language Institute
	Nederlandse Taalunie, Den Haag/The Hague Union for the Dutch Language
Norway	***Språkrådet***, Oslo The Language Council of Norway
Poland	***Rada Języka Polskiego przy Prezydium Polskiej Akademii Nauk***, Warszawa/Warsaw Council for the Polish Language
Romania	***Academia Română***, Bucureşti/Bucharest (Romanian Academy)
Slovakia	***Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied***, Bratislava Ludovit Stúr Institute of Linguistics, Slovak Academy of Sciences
Slovenia	***Služba za slovenski jezik, Ministrstvo za kulturo***, Ljubljana Slovenian Language Service – Ministry of Culture
	ZRC SAZU, Inštitut za slovenski jezik Frana Ramovša, Ljubljana (Fran Ramovš Institute of the Slovenian Language)
Sweden	***Språkrådet***, Stockholm Language Council of Sweden
	Svenska Akademien, Stockholm Swedish Academy
Serbia	***Институт за српски језик Српске академије наука и уметности***, Beograd/Belgrade Institute for Serbian Language of the Serbian Academy of Sciences and Arts

Switzerland	***Institute of Multilingualism***, Fribourg
Ukraine	***Секретаріат Уповноваженого із захисту державної мови***, Kyiv Secretariat of the State Language Protection Commissioner of Ukraine
United Kingdom	***The British Council***, London

Authors

Kozma Ahačič
Head of the Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Ljubljana, Slovenia

Júlia Choleva
Ludovit Stur Institute of Linguistics, Slovak Academy of Sciences

Radovan Fuchs
Minister of Science and Education of the Republic of Croatia

Federico Gaspari
Postdoctoral Researcher at the ADAPT Centre for Digital Content Technology in the School of Computing at Dublin City University, Ireland

Maria Giagkou
Computational Linguist at the Institute for Language and Speech Processing/ Athena RC, Greece

Nataša Gliha Komac
Deputy Head of the Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Ljubljana, Slovenia

Annika Grützner-Zahn
Junior Researcher in the Speech and Language Technology Lab, DFKI GmbH, Berlin, Germany

Jan Hajic
Director of LINDAT/CLARIAH-CZ Research Infrastructure and Deputy Director of the Institute of Formal and Applied Linguistics, CS School, Charles University, Prague, Czechia

Katrin Hallik
Institute of the Estonian Language (*until Dec 2021*)

Janoš Ježovnik
Deputy Head of the Fran Ramovš Institute of the Slovenian Language ZRC SAZU, Ljubljana, Slovenia

Paweł Kamocki
Legal expert, Leibniz-Institut für Deutsche Sprache, Mannheim, Germany

Sabine Kirchmeier
President of EFNIL

Nina Obuljen Koržinek
Minister of Culture and Media of the Republic of Croatia

Per Langgård
Oqaasileriffik/Language Secretariat of Greenland

Marek Łaziński
University of Warsaw, Council of Polish Language

Aino Piehl
Institute for the Languages of Finland

Stelios Piperidis
Senior Researcher, Head of the Natural Language Processing and Language Infrastructures (NLPLI) Department of ILSP/Athena RC, Greece

Georg Rehm
Principal Researcher and Research Fellow in the Speech and Language Technology Lab, DFKI GmbH, Berlin, Germany

Natalia Resende
Research Fellow in Natural Language Processing at the Centre for Applied Artificial Intelligence (CeADAR), School of Computer Science, University College Dublin, Ireland

German Rigau
Deputy Director of HiTZ, Basque Center for Language Technology, Donostia, Spain

Cecilia Robustelli
University of Modena and Reggio Emilia/Accademia della Crusca

Moritz Sommet
Institute of Multilingualism, Fribourg, Switzerland

Frieda Steurs
INT Leiden & KU Leuven

Elena Isabelle Tamba
Senior Researcher, Lexicographer at the Romanian Academy (*“A. Philippide” Institute of Romanian Philology*)

Trond Trosterud
UiT The Arctic University of Norway

Johan Van Hoorde
President of EFNIL 2018–2021

Andy Way
Deputy Director of the ADAPT Centre for Digital Content Technology in the School of Computing at Dublin City University, Ireland

Andreas Witt
Head of the Digital Linguistics Department, Leibniz-Institut für Deutsche Sprache, Mannheim, Germany