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.

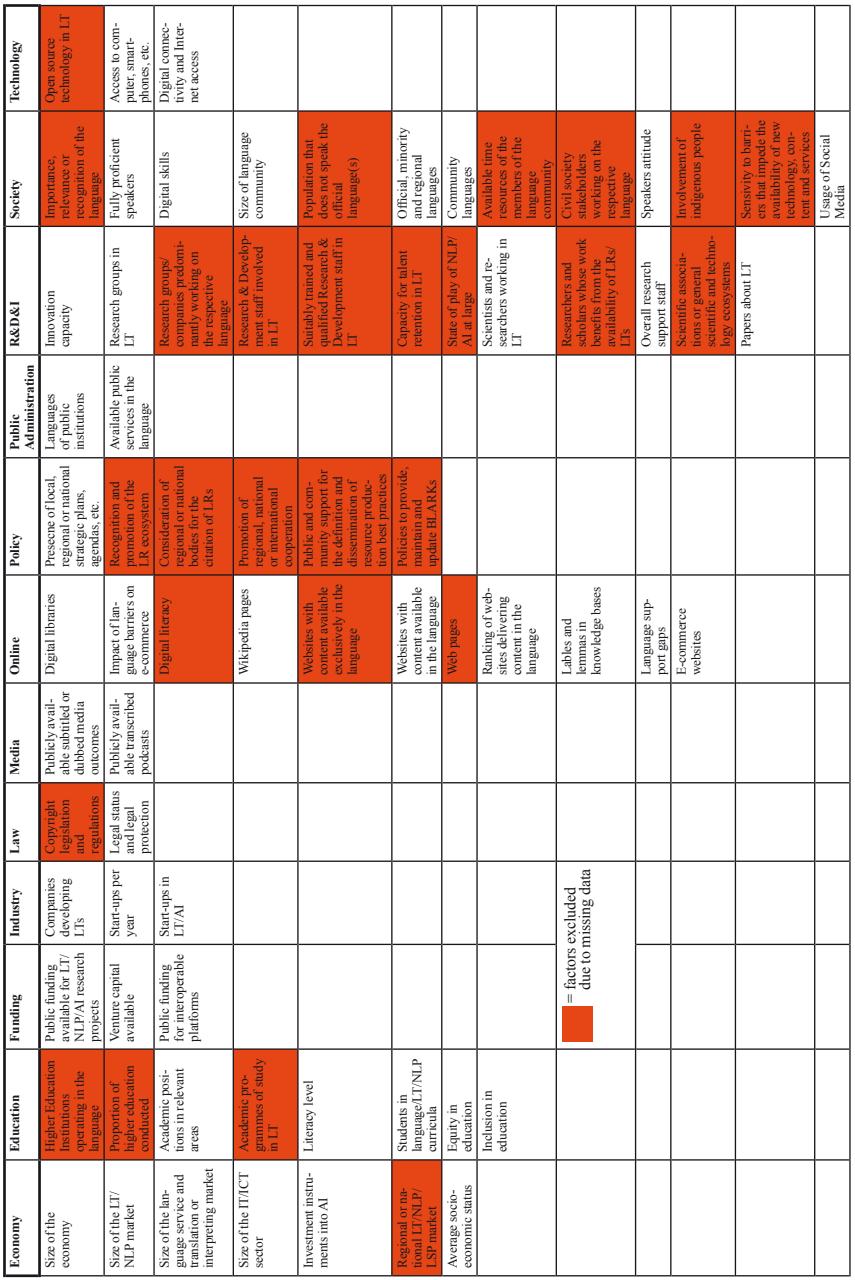

Fig. 1: Overview of the contextual factors

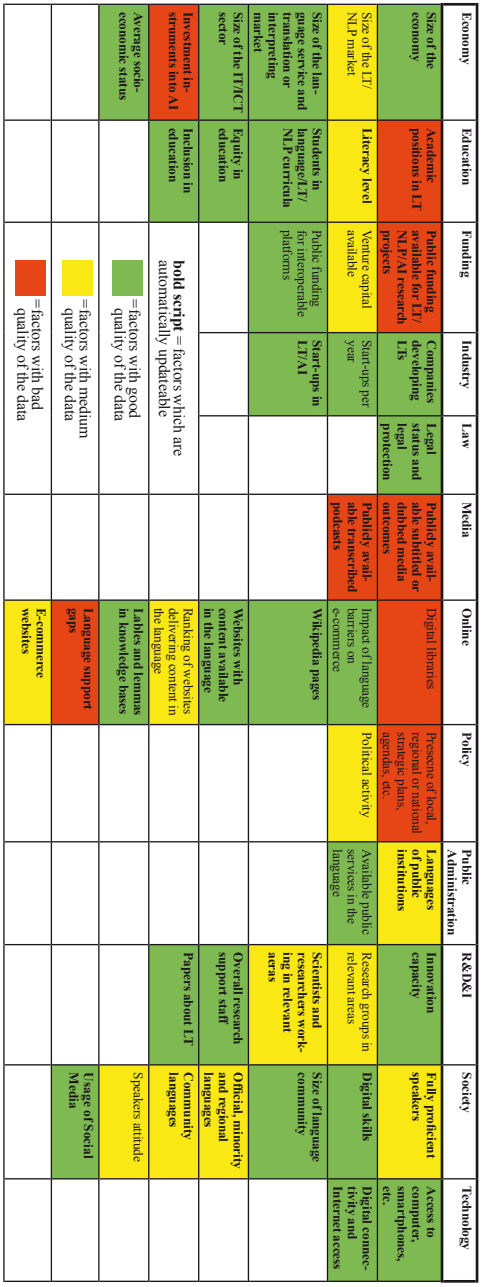

Fig. 2: Classification of the contextual factors

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.

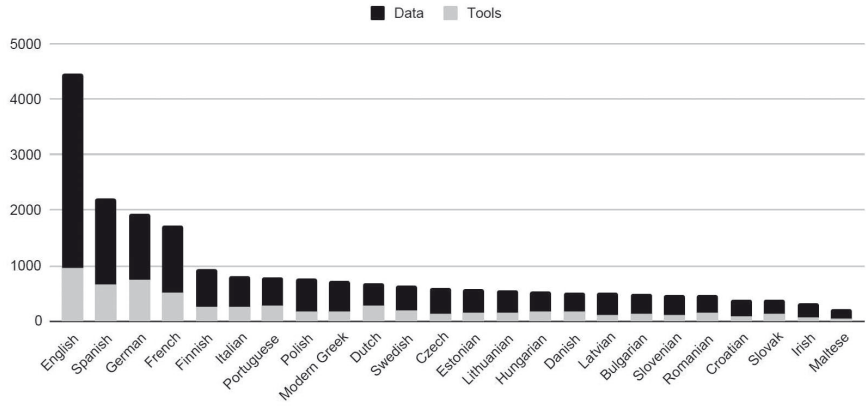

Fig. 3: Number of resources (data and tools) for the official EU languages

[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.

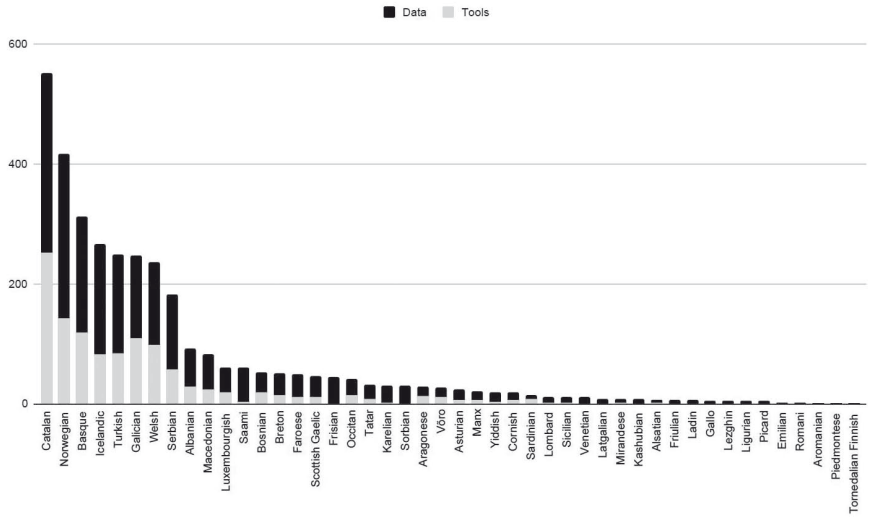

Fig. 4: Number of resources (data and tools) for various non-official EU languages

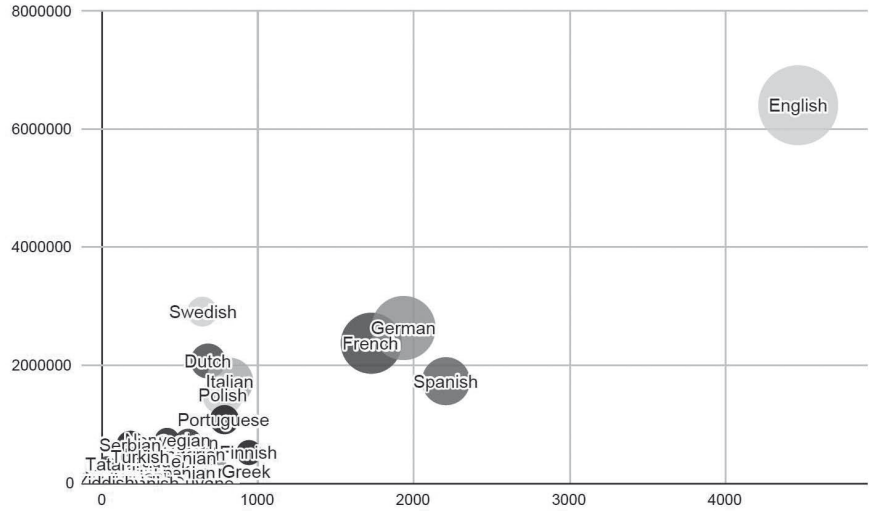

Fig. 5: Number of total resources in our collection vs. number of Wikipedia articles (the size of the circles represents the number of L1 and L2 speakers of the language in Europe)

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.

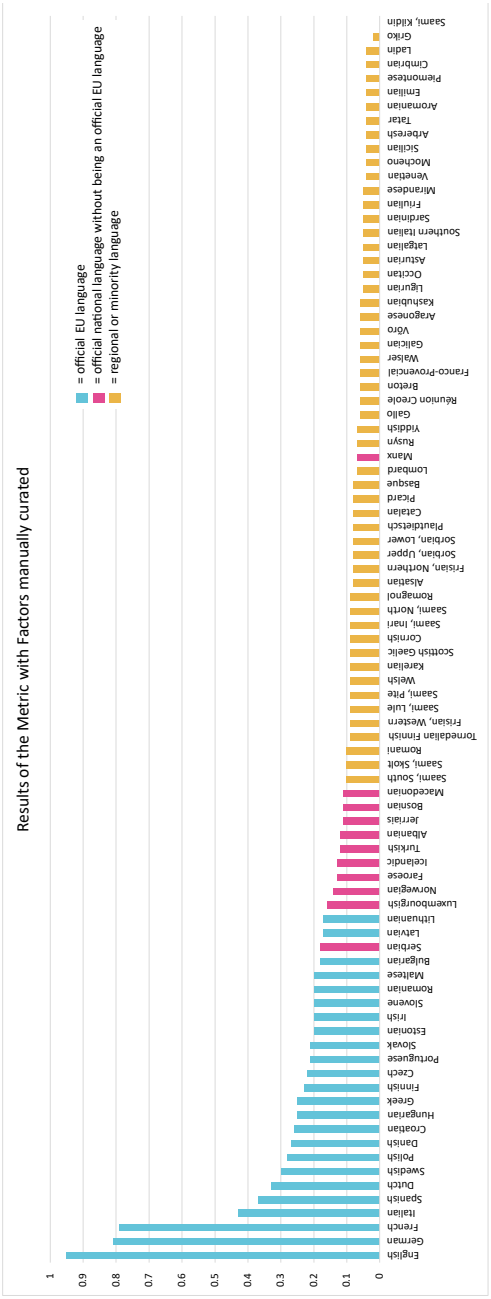

Fig. 6: Results of the 12 manually curated contextual factors

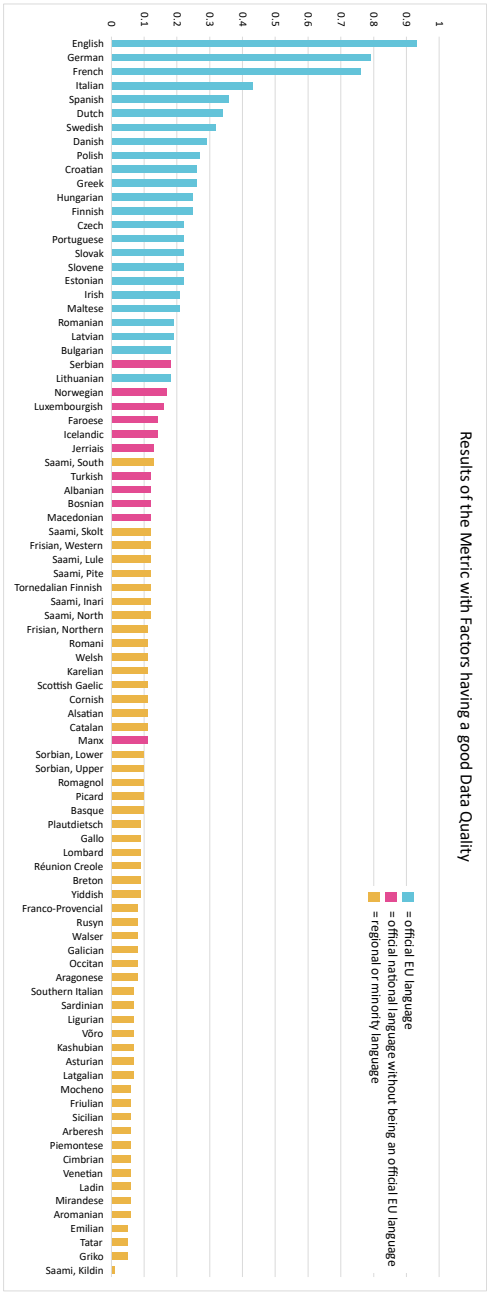

Fig. 7: Results of the 26 contextual factors with good quality of the data

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

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)

Bibliographical information

This text is part of the book:

Željko Jozić/Sabine Kirchmeier (eds.) (2022): The role of national language institutions in the digital age. Contributions to the EFNIL Conference 2021 in Cavtat. Budapest: Nyelvtudományi Kutatóközpont/Hungarian Research Centre for Linguistics.

This electronic PDF version of the text is accessible through the EFNIL website at:
http://www.efnil.org