Surveying the Technology Support of Languages

Annika Grützner-Zahn[1], Federico Gaspari[2], Maria Giagkou[3], Stefanie Hegele[1], Andy Way[2] and Georg Rehm[1]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI), Germany

[2]Dublin City University (DCU), Ireland

[3]R. C. "Athena", Institute for Language and Speech Processing (ILSP), Greece

Corresponding author: annika.gruetzner-zahn@dfki.de

Abstract

Many of the world's languages are left behind when it comes to Language Technology applications, since most of these are available only in a limited number of languages, creating a digital divide that affects millions of users worldwide. It is crucial, therefore, to monitor and quantify the progress of technology support for individual languages, which also enables comparisons across language communities. In this way, efforts can be directed towards reducing language barriers, promoting economic and social inclusion, and ensuring that all citizens can use their preferred language(s) in the digital age. This paper critically reviews and compares recent quantitative approaches to measuring technology support for languages. Despite using different approaches and methodologies, the findings of all analysed articles demonstrate the unequal distribution of technology support and emphasise the existence of a digital divide among languages.

Keywords: Less-Resourced/Endangered Languages, Language Resources and Technologies, Projects/Policy issues, Quantitative Evaluation Methodologies

1. Introduction

The field of Language Technology (LT) and Natural Language Processing (NLP) has seen huge progress recently. Cutting-edge technology is integrated into our daily lives more and more and used by hundreds of millions of people on a regular basis. Still, many applications are able to handle only a fraction of the approximately 7,000 languages, excluding a large number of potential users.

The ability to monitor the progress of technology support, and to make comparisons across languages, is essential to encourage further development for languages that are lagging behind the so-called 'major' languages. This is particularly pertinent in multilingual societies all over the world, where members of language communities poorly supported by technologies face economic, cultural and social disadvantages due to language barriers; without dedicated intervention, this situation is bound to become worse, eventually leading to digital language extinction of many low-resource languages, while speakers of the 'major' languages benefit from unprecedented, increasing international connectivity and all related advantages (Rehm and Uszkoreit, 2012; Kornai, 2013; Daly et al., 2023).

To obtain a realistic picture of the state of digital readiness of the world's languages, reliable indicators and agreed upon methods are needed to measure the level of technology support. The earliest investigations examining various languages in this regard found indicative evidence through qualitative methods, most notably the META-NET White Papers in 2012. Since then, and especially with the wide-scale adoption of neural methods, the field has made various breakthroughs.

Several quantitative approaches have recently been proposed to map what is happening in this space, usually with a view to tackling the observed inequalities by encouraging the LT/NLP community to address the identified gaps and shortcomings. However, these endeavours suffer from a lack of agreed upon methods of analysing the current state of affairs, so that comparisons across studies are essentially impossible. There are differences in the data analysed, a diverse range of concepts used, and different measures employed. This reflects not only a lack of agreement within the community, but also the possible different perspectives on the topic.

This paper analyses and compares recent approaches proposed to *quantitatively* measure the level of technology support of languages, based on a systematic review, following the PRISMA 2020 approach. This work aims at deepening our understanding of the many possible factors influencing the development of Language Resources and Technologies (LRTs) for different language communities and to provide a sound base for further examinations of consequences and solutions. Our results show that despite the heterogeneity of the approaches, the measures concerning LRTs can be differentiated between measures of quantity (how many LRTs are available?) and quality (how good are existing LRTs?). Most approaches consider socio-economic factors and examine de-

Proceedings of the Second TDLE Workshop @LREC-COLING-2024, pages 1–17

pendencies on research and the broader economy. All surveyed studies demonstrate an unequal distribution of technology support, proving the existence of a digital divide.

Section 2 of this paper summarises related work. Sections 3 and 4 describe the methodology and the articles that were analysed and compared as part of this systematic review. The results are presented in Section 5, while Section 6 discusses selected aspects in more detail and Section 7 concludes the paper.

2. Related Work

Language death is a threat to many small communities. Bromham et al. (2021) examine the effects of a range of demographic and socio-economic aspects on the use and status of the world's languages, and conclude that language diversity is endangered since half of the languages are at risk of extinction. This trend also applies in the digital sphere. The vitality analysis by Kornai (2013) shows that at least 2,500 languages are considered to be endangered.

The preservation of languages is a key goal of future LRT development (Rehm and Uszkoreit, 2012; Meighan, 2021; Daly et al., 2023). Many publications advocate for implementing ethical principles such as equity or equality, fairness, and diversity for languages in the digital realm (Carew et al., 2015; Soria, 2017; Bender and Friedman, 2018; Birhane, 2021; Choudhury and Deshpande, 2021; Ramesh et al., 2023; Rehm and Way, 2023). With regard to the development of LRTs, the focus should shift from optimising performance to a more holistic, human-centred perspective in order to serve all user communities (Ethayarajh and Jurafsky, 2020). Emerging technologies are used by all kinds of language communities around the globe, from small to large, as well as in traditionally oral contexts or deaf communities (Prasad et al., 2018; van Esch et al., 2019). Focusing on primarily oral languages, Bird (2022) argues that a shift is required which builds on the participation of local communities to identify new opportunities for LRTs in low-resource scenarios, abandoning the assumption that all languages can be served by the same technologies.

Krauwer (2003) provided one of the first calls for action towards more emphasis on under-resourced languages. Subsequent qualitative analyses of the technology support of languages continued to indicate a trend towards a digital divide between a few dominant and widely-used languages and many other languages, which are far less supported (or not at all), often spoken by smaller language communities (Yin et al., 2021; Khanuja et al., 2023). In addition to the linguistic bias, Helm et al. (2023) talk about a techno-linguistic or 'design' bias (Santy et al., 2023), which is expressed through the inability of systems to adapt to the knowledge systems of non-Western language communities.

More recent research focuses on "digitally-disadvantaged languages". Corresponding language communities encounter the following three main challenges because of missing LRT support: "1. gaps in equitable access; 2. digital tools that negatively impact the integrity of their languages, scripts and writing systems, and knowledge systems; and 3. vulnerability to harm through digital surveillance and under-moderation of language content" (Zaugg et al., 2022, pp. 2–3).

The articles analysed in this work provide data-driven evidence for the current situation. They indicate which languages need further financial support and research efforts to be able to mitigate current inequalities and biases within LRTs.

3. Methodology

In order to analyse and compare the approaches used to measure the level of technology support, we conducted a systematic review of articles, with a view to pointing out common and unique features as well as shortcomings that require further investigation. We followed the PRISMA 2020 statement, which includes a checklist of items for systematic reviews emphasising transparency, accuracy and completeness (Shamseer et al., 2015; Page et al., 2021a,b, 2022).

The earliest relevant article for this systematic comparative review, Joshi et al. (2020), seeks to gauge the technology support of individual languages from a quantitative perspective. Since then, further quantitative research has been published. We performed an extensive search in Google Scholar, going through a total of 419 citations (as of August 2023) for Joshi et al. (2020). Qualitative papers, literature reviews and benchmark evaluations were excluded, since our goal was exclusively the evaluation of novel quantitative approaches. To ensure that Joshi et al. (2020) was indeed the first publication of our interest, the publications referenced in the papers collected were checked for missing papers written in languages other than English, but none were found. These steps led to a set of nine papers (see Table 1).[1]

Five key criteria were defined: C1 *Research question* examines the different perspectives on the topic of technology support for languages; C2 *Scope* compares the number of languages,

[1] P7a and P7b are two complementary publications from the same project. While listed separately in Table 1, in the rest of the paper they are considered jointly.

ID	Reference	Year
P1	The State and Fate of Linguistic Diversity and Inclusion in the NLP World, P. Joshi et al.	2020
P2	Systematic Inequalities in [LT] Performance across the World's Languages, D. Blasi et al.	2022
P3	Dataset Geography: Mapping Language Data to Language Users, F. Faisal et al.	2022
P4	Some Languages are More Equal than Others: […], S. Ranathunga	2022
P5	Assessing Digital Language Support on a Global Scale, G. F. Simons et al.	2022
P6	Evaluating the Diversity, Equity and Inclusion of NLP Technology: […], S. Khanuja et al.	2022
P7a	Introducing the Digital Language Equality Metric: Technological Factors, F. Gaspari et al.	2022
P7b	Introducing the Digital Language Equality Metric: Contextual Factors, A. Grützner-Zahn et al.	2022
P8	Writing System and Speaker Metadata for 2,800+ Language Varieties, D. van Esch et al.	2022

Table 1: Scientific articles included in our literature review

geographical and LRT areas covered and number of measurements reported; C3 *Conceptualisation and quantification* analyses the ways in which the items to be measured are conceptualised and quantified; C4 *Combination of factors* considers how different factors were put in relation to one another; C5 *Results* compares the results of the paper with those of the other articles under review.

We manually extracted the relevant information from each article, looking separately at independent measures in each one, i. e., for measurements related to separate factors. While we succeeded in analysing each paper for each criterion, the diversity required different types of comparison, in particular, for "Conceptualisation and quanitification" and "Combination of factors".

In terms of possible reporting biases, the article selection process may have resulted in this survey missing those articles that do *not* cite Joshi et al. (2020), but that still conduct relevant research. This would also apply to articles published before 2020. Additionally, there are some borderline cases, such as the well-known article by Kornai (2013), which could not cite Joshi et al. (2020) and which does not fit our survey fully because its goal was to assess the vitality of languages (i. e., decline, endangerment, and eventually death). Another borderline case was the PhD thesis by Berment (2004) which provides a framework for the quantification of the computerisation of languages, but the data used for this quantification is based on expert knowledge and a possible application of the framework is only shown for one language. Furthermore, during the analysis of the results (Section 5), the European affiliation of the authors may have resulted in a stronger emphasis and focus on European languages in the assessment of how the results match our knowledge (and potentially expectations) about a language community and its situation.

4. Materials and Data Sources

Our survey includes the most significant contributions in this area. Eight papers were published in 2022 compared to only one in 2020, which points to a dynamic and fast-progressing area of work, whose importance is likely to increase.

Joshi et al. (2020) investigate the relation between the world's languages and resources as well as their coverage in NLP conferences: their analysis reveals a severe disparity across languages in terms of available data sets and coverage in research fora. Blasi et al. (2022) assess the global utility of LTs in relation to demographic or linguistic demand and analyse over 60,000 NLP conference papers, showing evidence for the unequal development of LTs across languages.

Faisal et al. (2022) argue that the availability of data is the decisive factor for the quality of NLP systems, and investigate the geographical representativeness of datasets, gauging to what extent they meet the needs of the language communities, exploring especially geographical and socioeconomic factors that may explain the dataset distributions. Ranathunga and de Silva (2022) look at linguistic disparity in NLP, using a categorisation of languages based on speaker population and vitality. They examine the distribution of LRs, the amount of research, inclusion in multilingual platforms and models among the categories and analyse the role of some contextual factors.

Simons et al. (2022) evaluate the level of digital language support through the extraction of language names from the websites of over 140 tools, and propose a categorisation of the languages based on the number of tools per LT area. Khanuja et al. (2023) focus on Indian languages and discuss an approach to evaluating NLP technologies based on diversity, equity and inclusion to quantify the diversity of the users they can serve. The method aims at addressing gaps in LRT provisioning related to societal wealth inequality.

Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022) present the Digital Language Equality metric, that quantifies the digital support of Europe's languages. Its technological factors measure the number of LRTs for each language within the European Language Grid (ELG, Labropoulou

et al., 2020; Rehm et al., 2021; Rehm, 2023),[2] which is assumed to be representative. The contextual factors reflect the broad socio-economic ecosystem of the languages by taking into account a set of indicators considered to be relevant.

Finally, van Esch et al. (2022) describe an open-source dataset which covers 2,800 languages and their writing system(s), along with estimates of the speaker populations. They analyse the distribution of languages and writing systems in language models (LMs), comparing it to the coverage of the respective language families in NLP research, hoping that this dataset will help develop NLP research for under-researched languages.

5. Results

5.1. C1 Research Question

The authors of all papers measured the technology support of languages independently, with no common objective or framework, and so decided to consider a number of different factors, which are difficult to directly compare or relate to each other. Already the specific research questions target different aspects and focus on subsets of areas, such as data availability or scientific output in NLP research. The approaches distinguish between the measurement of LRs, LTs or socio-economic factors assumed to have an impact on the development of LRTs. We identified 20 different research questions or aims (see Table 4 in the Appendix) evenly distributed across these three areas. One difference between the approaches is whether the goals are defined to measure either availability, coverage and quantity or performance and quality of the LRTs. Next to the more specific aims like "distribution of available data" (Joshi et al., 2020) or "platform interface availability" (Ranathunga and de Silva, 2022), some authors tried to approximate notions such as "inclusion" (Ranathunga and de Silva, 2022; Joshi et al., 2020) or "equity" (Khanuja et al., 2023) in the realm of LRTs. Just like Simons et al. (2022), Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022) define their own concepts. Interestingly, both approaches cover the largest number of LRT areas (see Section 5.2).

The analysis of scientific coverage of single languages as one of the most prevalent socio-economic factors is quite striking. The only other group of socio-economic factors directly mentioned as an object of measurement are geographic factors because of a specific focus on geographical representation (Faisal et al., 2022). In two papers, the socio-economic factors concerning the number of speakers (van Esch et al., 2022) and global demand (Blasi et al., 2022) are used as part of the ratios to LRT coverage or performance.

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

5.2. C2 Scope

The scope of an approach to measure the level of technology support can be thought of in different ways. The geographical coverage or number of languages investigated differs based on the focus of the research or data availability considered by each paper. Similarly, the numbers of languages addressed range from 15 to 7,829 (see Table 2). Faisal et al. (2022) propose a language-independent approach and do not state the number of languages in the datasets they analyse. Although six of the eight papers aim at surveying all languages of the world, only two actually cover more than 6,000 languages. The other approaches miss out on a huge number of languages that exist today in the world, despite the stated ambition to cover them all.

The influence of data availability is especially visible in the article by Blasi et al. (2022), where the number of languages under consideration varies dramatically depending on the task being evaluated, with values ranging between 15 and 630. A similar issue was observed for the socio-economic data (Grützner-Zahn and Rehm, 2022).

ID	Region	Number of Languages
P1	World	2,485
P2	World	Task-dependent (between 15 and 630)
P3	World	Not mentioned; language-independent
P4	World	6,420
P5	World	7,829
P6	India	22
P7	Europe	90
P8	World	2,800

Table 2: Targeted region and languages covered

The measurement approach can also reflect different aspects of technology support, such as quality of performance, quantity of LRTs or aspects such as efficiency. Table 3 shows that some papers focus on the creation of a single measurement concentrating on one aspect of technology support (e. g., Simons et al., 2022), while others establish a number of different, independent means (e. g., Ranathunga and de Silva, 2022, consider five). Those independent means can be mapped to different LRT areas (e. g., Ranathunga and de Silva, 2022), or a single area (Khanuja et al., 2023; van Esch et al., 2022); this raises the question of the extent to which the measurement of support for a language in a single LRT area can actually provide an accurate and reliable indication of the overall level of technology support that can also be compared across languages.

In contrast, a small number of approaches are based on broader coverage. Simons et al. (2022) and Gaspari et al. (2022) cover a relatively high number of LRT areas, nine and ten. Half of the papers cover only one or two LRT areas, which are taken to be good enough indicators to reflect the overall state of technology support. This applies to Joshi et al. (2020) and Khanuja et al. (2023), which also define in their aim of measurement general LT performance or broad social concepts such as "inclusion", "equity" or "diversity" (see Section 5.1). Seven out of eight papers (i. e., all papers except Simons et al., 2022) also consider socio-economic or contextual factors, such as scientific coverage or Gross Domestic Product (GDP).

ID	Number of Approaches	Number of LRT Areas	Socio-economic Indicators
P1	4	2	yes
P2	2	6	yes
P3	3	2	yes
P4	5	3	yes
P5	1	9	no
P6	3	1	yes
P7	2	10	yes
P8	2	1	yes

Table 3: Scope of measurement

Another way of approaching this aspect is the analysis based on the size of the datasets used in the papers. We did not include this dimension in our review because it would have required a lot of additional work with potentially little actual gain, particularly due to the difficulty of directly combining, and comparing across languages, different measures of the size of the datasets. In addition, the papers do not always provide all details concerning the size of the datasets they discuss, often only referring readers to other sources. Preliminary attempts to gather the details from other cited papers, archives and platforms required substantial effort without necessarily leading to conclusive results that could be confidently analysed or compared for the purposes of this survey.

5.3. C3 Conceptualisation and Quantification

For C3, the measurement approaches were divided into LRs, LTs and socio-economic indicators.

5.3.1. Language Resources

Five papers measure the availability of data for a language or the representation of language (community) features in the available data (see Table 6). Joshi et al. (2020) is the only one that takes both dimensions into account. Three papers use the raw counts of datasets with labelled and unlabelled data in different repositories to approximate the availability of language data. Joshi et al. (2020) add the question of the distribution of these data resources. The distribution is exemplified through the classification of languages and further analysis of size. Ranathunga and de Silva (2022) focus solely on the coverage of languages. They reuse the approach from Joshi et al. (2020), but add another repository, Hugging Face, to the repositories used by Joshi et al. (2020), namely LDC and ELRA. A weighting of datasets based on their features is introduced by Gaspari et al. (2022) who use as data source the ELG which harvests several major repositories such as Zenodo and CLARIN. Gaspari et al. (2022) mention the problem of dataset size, which is difficult to measure because of missing data and incompatible descriptions and values, while recognising that it would be desirable to include this information.

Joshi et al. (2020), Faisal et al. (2022), and van Esch et al. (2022) analyse the representation of language features or local knowledge in the datasets. All three focus on different aspects of language representation which reflect the layers of diversity between languages and their communities. Joshi et al. (2020) examine which language features are not represented in the datasets. This typological conceptualisation is motivated by transfer learning and the idea that less-resourced languages can reach a better level of support if their features are covered in LMs. van Esch et al. (2022) also concentrate on a language's writing system. The share of a writing system in the vocabulary of multilingual models is calculated, from which the authors induce the representation in NLP. Both conceptualisations are motivated by language modelling (either through its learning mechanisms or the training data) and directly compare the languages with each other based on the chosen feature. Faisal et al. (2022) deviate from this through a focus on local knowledge contained in language data. The number of local entities in the dataset reflects the distance of the dataset from the users and their needs through language and LT-task-independent means. Nonetheless, only datasets designed for Named-Entity-Recognition (NER) and Question Answering (QA) are analysed in the paper. While Joshi et al. (2020) use a dataset that is independent of the LT area, Faisal et al. (2022) and van Esch et al. (2022) use LT-specific datasets and deduce the general concept of representation within NLP.

5.3.2. Language Technologies

Four papers include measures of LT performance, while three papers contain measures of LT availability, but none combine both perspectives (see

Table 7). The papers mainly use known performance measures for certain NLP tasks, such as reused Natural Language Inference (NLI) error rates (Artetxe and Schwenk, 2019; Joshi et al., 2020). Faisal et al. (2022) calculate F1 scores in the context of QA and the influence of geographical representation in the training datasets. Blasi et al. (2022) introduce a measure of utility which is quantified as performance divided by a theoretical maximum performance. The utility per LT area is added up to reach an overall score per language. Overall, the possibility to summarise LT performance of different LT tasks, despite the use of different performance measures, gives a broader picture than the approaches covering single LT areas. Khanuja et al. (2023) use different means to measure the performance of LMs. They reuse the utility metric from Blasi et al. (2022), but extend it through projected performance estimates for languages without available test data (otherwise set to 0) based on the performance of languages from the same family and the availability of unlabelled data. This extension is motivated by transfer learning and the possible performance increase, if language features are learnt from another language. Since the utility measure assigns the same scores to the languages covered by one model, despite different performance on different languages, another measure is proposed to account for the equity in model performance, namely the Gini-coefficient, which measures the inequality within a distribution (model performance on different languages). A third measure reflects inclusion through model efficiency concerning the use of computational resources. This last measure of efficiency shows that other methods of evaluating the "quality" may be of importance, even though they are considered by only one paper.

Similar to the LR availability measures, the LT availability measures are quantified as counts of services available for the languages. Ranathunga and de Silva (2022) collected the languages in which Google Translate and Facebook are available. Similarly, the languages covered by mBERT and XLM-R are counted, to provide an approximation for general model coverage. In the article by Simons et al. (2022), Digital Language Support is conceptualised as a scale with a strict support-level hierarchy, in which each level is quantified through the number of popular tools available for each LT area. For each LT area, the ten most popular tools globally and the five most popular tools of each of the ten largest countries in terms of population were selected. An approach to combining several LT areas into one metric was also chosen by Gaspari et al. (2022), based on the number of available LTs in ELG per language. Again, the authors included a weighting mechanism into the calculation of a score representing LT support, assuming that some LTs are more demanding to develop than others. While Ranathunga and de Silva (2022) analyse only two platforms and two models, Simons et al. (2022) and Gaspari et al. (2022) quantify the availability of LTs in a broader and more comprehensive way.

5.3.3. Socio-economic Indicators

The socio-economic factors are often approximated with indicators of scientific output or inclusion. A typical quantification of scientific output is the number of publications concerning a particular language (Joshi et al., 2020; Ranathunga and de Silva, 2022; Grützner-Zahn and Rehm, 2022; van Esch et al., 2022). Alternatively to the use of plain figures (Ranathunga and de Silva, 2022; van Esch et al., 2022), Joshi et al. (2020) use language occurrence entropy as a proxy for language diversity in NLP conferences. Another perspective is the use of reputation quantified as the number of citations (Blasi et al., 2022) or the prediction of proximity through embeddings in which authors, languages and conferences serve as entities and the title and abstract as context (Joshi et al., 2020).

In all cases, the economic situation is quantified with GDP, while the size of a language community is defined as the number of speakers, although often the information about how people qualify as speakers (acknowledged as a difficult issue) is missing. Blasi et al. (2022) quantify demographic demand using also the number of speakers, while contrasting it to linguistic demand. Faisal et al. (2022) introduce a geographical distance, reflecting the distance between user and producer based on entities in a dataset, and country size. The situation of a language and its speakers can be measured by a range of factors. Most authors focus on just a few, perceived as most relevant for the development of LRTs. Grützner-Zahn and Rehm (2022) present the only approach trying to combine socio-economic factors from different areas (such as science, education, economy, etc.) to paint an overall picture on a single scale of the specific contexts of Europe's languages as part of the Digital Language Equality concept.

5.4. C4 Combination of Factors

The approaches presented in the eight papers differ substantially in terms of how their indicators contribute to the bigger picture. While some only represent single indicators and their results, others create a metric in which the indicators are combined to a ratio (all except Ranathunga and de Silva, 2022). Some approaches measure the relation between two factors through co-occurrence or correlation measures (see Table 10

in the Appendix for the relevant details). While Blasi et al. (2022), Simons et al. (2022) and Gaspari et al. (2022) along with Grützner-Zahn and Rehm (2022) combine indicators of different types representing different LRT areas or even socio-economic factors into one overall score as a result, Joshi et al. (2020), Faisal et al. (2022) and van Esch et al. (2022) create ratios within one area of application. Khanuja et al. (2023) aim to measure the three concepts of "diversity", "equity" and "inclusion" creating ratios combining different aspects of model performance.

Joshi et al. (2020) and Simons et al. (2022) assign languages to classes. While in Joshi et al. (2020) the class of the language is derived from the data availability measure, Simons et al. (2022) distinguish classes of digital language support based on the coverage of available LTs. The step of including a classification on top of the scores is left out by Blasi et al. (2022) and Gaspari et al. (2022) along with Grützner-Zahn and Rehm (2022), although both approaches also result in overall scores per language.

The papers examining the relation between two factors use either basic occurrence measures searching for patterns or outliers (Joshi et al., 2020; Ranathunga and de Silva, 2022; van Esch et al., 2022) (see Table 9 in the Appendix) or correlation measures (Blasi et al., 2022; Faisal et al., 2022; Ranathunga and de Silva, 2022). Ranathunga and de Silva (2022) use the occurrence measures to identify the outlier, and analyse it through an additional correlation measure. In contrast, Blasi et al. (2022) and Faisal et al. (2022) use different kinds of correlation measures to examine which socio-economic factor (e. g., GDP or number of speakers) best predicts the result, such as the number of papers published on a language or the representation of language communities in a dataset.

5.5. C5 Results

All papers identified a digital divide between a few dominant languages and a majority of low-resourced languages. Not surprisingly, English is always, by far, the best supported language in all LRT areas when languages are compared directly, usually followed by Spanish, German and French (Joshi et al., 2020; Ranathunga and de Silva, 2022; Gaspari et al., 2022; Grützner-Zahn and Rehm, 2022), three official European Union languages with large numbers of speakers. Ranathunga and de Silva (2022) detect a bias towards Indo-European languages spoken in Europe and institutional languages[3] with large speaker populations. Still, even within Europe a huge imbalance was identified by Gaspari et al. (2022) and Grützner-Zahn and Rehm (2022). Regional and minority languages (RMLs) have mostly been ignored (with a few exceptions, such as Basque, van Esch et al., 2022). The authors conclude that much additional effort is needed to bridge the gap, although even most official languages lag way behind the 'major' languages mentioned.

Concerning data availability, more than half (Simons et al., 2022) or even 80% (Joshi et al., 2020) of the languages lack enough data to develop LT applications. Additionally, the size of the dataset decreases with the language class, meaning that even those datasets available for low-resourced languages are substantially smaller (Ranathunga and de Silva, 2022). Task-oriented datasets were found to have the highest counts for popular NLP tasks on large institutional languages, such as Machine Translation (MT) (Ranathunga and de Silva, 2022).

Most datasets exhibit biases towards the global west (Faisal et al., 2022) or linguistic feature representation of Indo-European or large official languages (Joshi et al., 2020; van Esch et al., 2022). Faisal et al. (2022) claim an unrepresentative number of entities in the data, but also find differences between datasets, such as MasakhaNER and Natural Questions from Google, which include a high proportion of entities from all over the world. Imbalances concerning linguistic features were detected to the extent that usually 2.86 categories per language feature are not represented in language data (Joshi et al., 2020). Combined with a measure showing higher error rates for languages containing these features, the results highlight the importance of language representation in data. Similarly, Faisal et al. (2022) show a decrease in performance on QA, if local knowledge is not included in training datasets.

For the development of LTs, Simons et al. (2022) find a correlation to higher categories of digital language support, implying that a strong basis of LRs and basic LT tools seems to be needed for the development of the higher categories, such as virtual assistants. Additionally, the results of the LT areas by Blasi et al. (2022) show that the majority of morphological or syntactic tools perform quite well if enough data is available. For MT and Text-to-Speech, the performance differs substantially among the languages with medium technology support. For complex tasks, such as NLI and QA, most systems perform poorly except for a few large official languages for which performance allows for actual use in operational settings. The performance of multilingual LMs shows a huge im-

[3] Ranathunga and de Silva (2022) introduce the term "institutional languages" as a class of languages. The term is used in this paper to avoid confusing the discussion of their results.

balance even for the best models, although region-specific tuning can counteract the limited transfer between languages in a multilingual model to a certain extent (Khanuja et al., 2023). Similar results were shown by Simons et al. (2022) and Gaspari et al. (2022). While basic LTs are available for a considerable number of languages, the number quickly decreases for less-resourced languages as the complexity of the tools grows.

The analyses of the socio-economic factors show a similar pattern. The scores for contextual factors (Grützner-Zahn and Rehm, 2022) describe an uneven distribution towards large official languages in Europe, while RMLs receive little attention from the economy, politics, etc. The results make national and regional efforts towards the support of RMLs visible, e. g., the co-official languages in Spain achieve relatively high scores compared to RMLs with similar numbers of speakers elsewhere. Correlation measures give indicative evidence that the GDP and/or geographical distance are the two socio-economic factors that best predict the amount of NLP research and development (Blasi et al., 2022; Faisal et al., 2022; Ranathunga and de Silva, 2022). The best predictor for geographical representation constitutes a ratio of the two factors, reflecting that potentially many socio-economic factors have an impact on LRT development (Faisal et al., 2022), as considered by Grützner-Zahn and Rehm (2022) in which a ratio of socio-economic factors is calculated. Although Blasi et al. (2022) and Faisal et al. (2022) show that the inclusion of speaker population causes the prediction to deteriorate, the number of speakers is considered by most papers analysing socio-economic factors (Blasi et al., 2022; Faisal et al., 2022; Khanuja et al., 2023; Grützner-Zahn and Rehm, 2022; van Esch et al., 2022).

The focus on NLP research in most papers shows a more fine-grained picture. Large European languages are considerably more often the subject of research, and more popular languages are in turn propagated more, making the existing imbalance even worse (Joshi et al., 2020). Additionally, the number of languages addressed in a publication does not predict the number of citations a paper is going to receive, i. e., there is no incentive for researchers to address a larger number of languages. Still, focused research communities have been detected for some non-European or non-official languages, such as Japanese, Turkish, Inuktitut, Hawaiian, etc. (Joshi et al., 2020; Blasi et al., 2022), and even among those languages with large speaker populations, some are underrepresented (van Esch et al., 2022), showing that concentrated efforts are picked up by quantitative measures, and that it is not all about size.

Some authors classify languages based on the resulting scores. However, classifications create hard boundaries, introducing a distinction between languages which might otherwise be thought of as having similar levels of support, e. g., Simons et al. (2022) assign Hungarian the class "Thriving", while Latvian is "Vital". In Gaspari et al. (2022), though, Hungarian and Latvian achieve similar scores. In contrast, some languages which appear to have different levels of support are grouped together. Simons et al. (2022) classify Latvian, Occitan and Yiddish as "Vital", but they obtain very different scores in Gaspari et al. (2022). Comparing size proportions between the approaches using a taxonomy, Joshi et al. (2020) group 88% of the languages in the lowest class, while 50% of the languages constitute the lowest class in Simons et al. (2022). Overall, this paints different pictures. Moreover, Ramesh et al. (2023) show that adding another data source changes the classification for 87 languages based on data availability. They conclude that single classifications should be avoided.

6. Discussion

All papers use very diverse approaches to measure the level of technology support of languages. Some authors chose to use notions from other fields such as "demand" and "utility" (Blasi et al., 2022) or "inclusivity", "equity" and "accessibility" (Khanuja et al., 2023). These are used in different ways and can be ambiguous if not properly defined and operationalised with respect to the languages under consideration. For instance, the definition of demand depends on the background, e. g., in economics it is viewed as the need of goods by consumers and may or may not include the will to pay depending on the use case (Rinkinen et al., 2020). In Blasi et al. (2022), demand is conceptualised from two angles, resulting in different outcomes for the metric. Demographically, demand was quantified through the number of speakers, but who exactly counts as a speaker and which (type of) speaker needs which (kind of) LT can be debated. Further explorations of how different quantification of demand may influence the results of the metric would be desirable to better assess its impact and to argue for a specific way of quantifying demand. The same applies to the other concepts mentioned above.

When analysing large datasets, biases can have an impact on different levels of the study: • Dataset assessment: analysis of biases in a dataset or study reused; • Study assessment: analysis of what kind of biases may be introduced through the choice of quantification, methodology, etc.; • Outcome-level assessment: analysis of biases in the results; • Reporting bias assessment: detecting whether all relevant results are made avail-

able. Not all levels need to be analysed in all studies, but dataset assessment is applicable to all studies, because they all reuse data. Only Ranathunga and de Silva (2022) describe possible biases through their chosen methodology and data in the appendix.

One possible source for bias has to do with the Bender rule (Bender, 2019). Many authors do not explicitly mention the language(s) covered, which is why figures about number of publications per language inevitably miss relevant publications. Another question has to do with how a speaker of a language is defined. Are L2 speakers considered? And if so, how reliable are the figures? A closer look into Ethnologue shows that many figures concerning the number of speakers are outdated, only contain L1 speakers or derive the number of speakers from the citizenship of the individuals, which distorts the numbers, especially in certain countries and regions. The question of which tools to include when approximating the technology support of a language can also introduce biases. Meighan (2021) and Bird (2022) show that some smaller language communities develop their own LRTs. Criteria such as tool popularity miss these developments and may fail to detect smaller advancements, that however may be significant for the language communities in question.

In Section 5.5, only parts of the results of the eight papers could be covered since not all findings were published; only Faisal et al. (2022), Gaspari et al. (2022), Grützner-Zahn and Rehm (2022) and van Esch et al. (2022) published all results. Simons et al. (2022) published 10% of their results which facilitates traceability, but does not allow extensive comparisons with other research. Joshi et al. (2020), Blasi et al. (2022), Ranathunga and de Silva (2022) and Khanuja et al. (2023) do not provide their full results or datasets. Thus, only the results described in these papers could be included in this survey.

7. Conclusions and Future Work

The systematic comparison of the eight papers under examination has shown that despite the heterogeneous approaches and differences on all levels of analysis, the results clearly indicate a very uneven distribution of LRTs between large, official, mostly Indo-European languages and essentially all other languages. The papers highlight different aspects, such as the output of focused research communities on specific languages or the influence of local knowledge on the performance of LMs. Combining all results to assemble a bigger picture reveals the many dependencies between all areas of LRTs and socio-economic factors. Efforts are needed on all levels, starting with data collection, for at least half of the world's languages.

Future work needs to examine how to standardise and measure the size of LRs and, similarly, the scope of LT applications. Another open issue is the actual quality of LRTs. Moreover, biases need to be further analysed, especially concerning their influence on the quantitative measures. All approaches we analysed cover only parts of the relevant measures, which is why the development of a measure accounting for all qualitative and quantitative perspectives, and covering all LRT areas would be an important step forward. Based on such an all-encompassing approach, further steps towards evaluation and the examination of possible solutions could be conducted.

Ethical Statement

The research described in this paper does not require typical ethical considerations. Having said that, when considering ethical aspects, the authors believe that analysing equality or equity, fairness and diversity within the area of Language Resources and Technologies is a timely and crucial topic that, on a general level, deserves much more attention in our research field. Additionally, the authors are all located in Europe which may have resulted in a focus on European languages in the analysis of the results because the authors are more familiar with them.

Acknowledgements

The European Language Equality project has received funding from the European Union under the grant agreements no. LC-01641480 – 101018166 (ELE) and LC-01884166 – 101075356 (ELE2), see https://european-language-equality.eu.

Bibliographical References

Mikel Artetxe and Holger Schwenk. 2019. Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Emily Bender. 2019. The #BenderRule: On Naming the Languages We Study and Why It Matters. *The Gradient*.

Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Vincent Berment. 2004. *Méthodes pour informatiser les langues et les groupes de langues " peu dotées "*. Ph.D. thesis, Université Joseph-Fourier - Grenoble I.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7817–7829, Dublin, Ireland. Association for Computational Linguistics.

Abeba Birhane. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2021. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Margaret Carew, Jennifer Green, Inge Kral, Rachel Nordlinger, and Ruth Singer. 2015. Getting in Touch: Language and Digital Inclusion in Australian Indigenous Communities. *Language Documentation & Conservation*, (9):307 – 323.

Monojit Choudhury and Amit Deshpande. 2021. How Linguistically Fair Are Multilingual Pre-Trained Language Models? *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12710–12718.

Emma Daly, Jane Dunne, Federico Gaspari, Teresa Lynn, Natalia Resende, Andy Way, Maria Giagkou, Stelios Piperidis, Tereza Vojtěchová¡, Jan Hajič, Annika Grützner-Zahn, Stefanie Hegele, Katrin Marheinecke, and Georg Rehm. 2023. Results of the Forward-looking Community-wide Consultation. In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 245–262. Springer, Cham, Switzerland.

Kawin Ethayarajh and Dan Jurafsky. 2020. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online. Association for Computational Linguistics.

Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. 2022. Dataset Geography: Mapping Language Data to Language Users. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3381–3411, Dublin, Ireland. Association for Computational Linguistics.

Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. 2022. Introducing the Digital Language Equality Metric: Technological Factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 1–12, Marseille, France. European Language Resources Association.

Annika Grützner-Zahn and Georg Rehm. 2022. Introducing the Digital Language Equality Metric: Contextual Factors. In *Proceedings of The Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 13–26, Marseille, France. European Language Resources Association.

Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2023. Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice. Publisher: arXiv Version Number: 1.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293, Online. Association for Computational Linguistics.

Simran Khanuja, Sebastian Ruder, and Partha Talukdar. 2023. Evaluating the diversity, equity, and inclusion of NLP technology: A case study for Indian languages. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1763–1777, Dubrovnik, Croatia. Association for Computational Linguistics.

András Kornai. 2013. Digital Language Death. *PLoS ONE*, 8(10):e77056.

Steven Krauwer. 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In *Proceedings of SPECOM 2003*, Moscow, Russia. Moscow State Linguistic University.

Penny Labropoulou, Katerina Gkirtzou, Maria Gavriilidou, Miltos Deligiannis, Dimitris Galanis, Stelios Piperidis, Georg Rehm, Maria Berger, Valérie Mapelli, Michael Rigault, Victoria Arranz, Khalid Choukri, Gerhard Backfried, José Manuel Gómez Pérez, and Andres Garcia-Silva. 2020. Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3421–3430, Marseille, France. European Language Resources Association (ELRA).

Paul J Meighan. 2021. Decolonizing the digital landscape: the role of technology in Indigenous language revitalization. *AlterNative: An International Journal of Indigenous Peoples*, 17(3):397–405.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. 2021a. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, page n71.

Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, and David Moher. 2021b. Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, 134:103–112.

Matthew J. Page, David Moher, and Joanne E. McKenzie. 2022. Introduction to PRISMA 2020 and implications for research synthesis methodologists. *Research Synthesis Methods*, 13(2):156–163.

Manasa Prasad, Theresa Breiner, and Daan Van Esch. 2018. Mining Training Data for Language Modeling Across the World's Languages. In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 61–65. ISCA.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023. Fairness in Language Models Beyond English: Gaps and Challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Georg Rehm, editor. 2023. *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer, Cham, Switzerland.

Georg Rehm, Stelios Piperidis, Kalina Bontcheva, Jan Hajic, Victoria Arranz, Andrejs Vasiļjevs, Gerhard Backfried, José Manuel Gómez Pérez, Ulrich Germann, Rémi Calizzano, Nils Feldhus, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Julian Moreno-Schneider, Dimitris Galanis, Penny Labropoulou, Miltos Deligiannis, Katerina Gkirtzou, Athanasia Kolovou, Dimitris Gkoumas, Leon Voukoutis, Ian Roberts, Jana Hamrlová, Dusan Varis, Lukáš Kačena, Khalid Choukri, Valérie Mapelli, Mickaël Rigault, Jūlija Meļņika, Miro Janosik, Katja Prinz, Andres Garcia-Silva, Cristian Berrio, Ondrej Klejch, and Steve Renals. 2021. European Language Grid: A Joint Platform for the European Language Technology Community. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations (EACL 2021)*, pages 221–230, Kyiv, Ukraine. Association for Computational Linguistics (ACL).

Georg Rehm and Hans Uszkoreit, editors. 2012. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages. Springer, Heidelberg etc.

Georg Rehm and Andy Way, editors. 2023. *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer International Publishing, Cham.

Jenny Rinkinen, Elizabeth Shove, and Greg Marsden. 2020. *Conceptualising Demand: A Distinctive Approach to Consumption and Practice*, 1 edition. Routledge, Abingdon, Oxon ; New York, NY : Routledge, 2021.

Sebastin Santy, Jenny T. Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. Publisher: arXiv Version Number: 1.

L. Shamseer, D. Moher, M. Clarke, D. Ghersi, A. Liberati, M. Petticrew, P. Shekelle, L. A. Stewart, and the PRISMA-P Group. 2015. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, 349(jan02 1):g7647–g7647.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Claudia Soria. 2017. What is Digital Language Diversity and why should we care? In *Digital Media and Language Revitalisation*, number 4 in Linguapax Review 2016, pages 13–.28.

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing System and Speaker Metadata for 2,800+ Language Varieties. In *Proceedings of the Language Resources and Evaluation Conference*, pages 5035–5046, Marseille, France. European Language Resources Association.

Daan van Esch, Elnaz Sarbar, Tamar Lucassen, Jeremy O'Brien, Theresa Breiner, Manasa Prasad, Evan Crew, Chieu Nguyen, and Françoise Beaufays. 2019. Writing Across the World's Languages: Deep Internationalization for Gboard, the Google Keyboard. Publisher: arXiv Version Number: 1.

Kayo Yin, Amit Moryossef, Julie Hochgesang, Yoav Goldberg, and Malihe Alikhani. 2021. Including Signed Languages in Natural Language Processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7347–7360, Online. Association for Computational Linguistics.

Isabelle A. Zaugg, Anushah Hossain, and Brendan Molloy. 2022. Digitally-disadvantaged languages. *Internet Policy Review*, 11(2).

A. Appendix

ID	Aim of Measurement
P1	Distribution of available data over languages
	Typological features of languages represented in data and influence of missing representation on LT performance
	Language diversity and inclusion of NLP conferences
	Closeness of authors, conferences and languages
P2	To what degree is the global demand for LT met?
	Correlation of scientific production in NLP
P3	Geographical representativeness of NLP datasets
	Socio-economic correlates
	Geographical breakdown of models performance
P4	Annotated data availability
	Platform interface availability
	Model coverage
	Amount of research conducted for the languages
P5	Digital language support
P6	Diversity
	Equity
	Inclusion
P7	Digital language equality
P8	Representation of writing systems in NLP compared to their speaker numbers
	Distribution of published works that reference the languages

Table 4: Research questions or aims

ID	LRT Areas Covered
P1	Data; Natural Language Inference
P2	Morphological Inflection; Syntactic Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference
P3	Data; Language Modelling
P4	Data; Human-Computer-Interaction; Language Modelling
P5	Data; Encoding; Morphological Inflection; Syntactical Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference; Human-Computer-Interaction
P6	Language Modelling
P7	Data; Encoding; Morphological Inflection; Syntactical Parsing; Text-to-Speech; Machine Translation; Question Answering; Natural Language Inference; Human-Computer-Interaction; Language Modelling
P8	Data

Table 5: LRT areas covered

ID	Concept	Conceptualisation	Values Used
P1	Distribution of available data	Creation of language taxonomy based on available data	labelled data unlabelled data
	Representation of typological features	Number of typological features not represented in languages often covered by LRs	Language features language taxonomy
P3	Geographical representativeness	Occurrence of entities associated with countries	entities and geographical connection languages and geographical connection
P4	Coverage by resources	Resource containing data in respective language	languages covered by selected resources
P7	Digital Language Equality	LRs contained in ELG (Including LTs and Contextual Factors)	Resource type Subclass Linguality type Media type Annotation type Domain Conditions of use
P8	Representation of writing systems	Scripts represented in model vocabularies	proportional share of words in script

Table 6: Conceptualisation of approaches covering LRs

ID	Concept	Conceptualisation	Values Used
P1	LT performance	error rates	Reuse of error rates from Artetxe and Schwenk (2019)
P2	Utility	sum of proportions of language performance to theoretical maximum performance per task	actual language performance theoretical maximum performance
P3	Model performance	comparison of model accuracy on question-answering test dataset	f1-scores
P4	Platform interface availability	languages covered by platform	languages covered by platform
	Model coverage	languages covered by model	languages covered by model
P5	Digital language support	support of languages by specific software products covering digital language support categories	number of tools digital language support categories
P6	Diversity	Reuse of conceptualisation from paper 2	
	Equity	Gini-coefficient for LT performance	LT Performance
	Inclusion	model efficiency	Throughput (= number of instances it can process per second on a CPU) Memory saved (= size of model as a measure how expensive a model is to use in practise) benefit (= model performance)
P7	Digital Language Equality	LTs contained in ELG (Including LRs and Contextual Factors)	Language dependence Input type Output type Function type Domain Conditions of use

Table 7: Conceptualisation of approaches covering LTs

ID	Concept	Conceptualisation	Values Used
P1	Language diversity and inclusion in conferences	Language occurrence entropy	Number of conference papers mentioning respective language year
	Closeness	Prediction of entity embeddings based on the context	entities: author, language, conference
P2	demographic demand	Size of language community	Number of speakers
	linguistic demand	Always highest value	1
	Reputation gain in research	number of citations	Number of citations
	scientific production	Publications in NLP	Number of NLP conference papers
	economic gain	GDP	approximate GDP of number of users
P3	size of community	population of country	population of country
	economic gain	GDP	GDP of country GDP per capita of country
	Size of country	landmass	landmass
	Distance between user and dataset	geographical distance between entities referenced in dataset and respective language community	location of entities location of language community
P4	economic gain	GDP	GDP of country
	size of language community	population size	number of speakers
P7	Digital Language Equality	Contextual Factors (Including LRTs/ Technological Factors):	
		Size of economy	Size of economy, Size of the ICT sector
		Education	Students in LT/language, Inclusion in education
		Industry	Companies developing LTs
		Law	Legal status and legal protection
		Online	Wikipedia pages
		R & D & I	Innovation Capacity, Number of papers
		Society	Size of language community, Usage of social media
		Technology	Digital connectivity, internet access

Table 8: Conceptualisation of socio-economic indicators

ID	Factor 1	Factor 2	Method
P1	Error rates	Representation of typological features	Mapping features not included in datasets and their error rates
P2	Number of normalised citations	Number of languages covered	correlation calculated based on Bayesian generalised additive mixed effects models
	GDP	Numbers of papers published	regression calculated based on Bayesian generalised additive mixed effects models
	Number of speakers	Numbers of papers published	regression calculated based on Bayesian generalised additive mixed effects models
P3	geographical distribution	country population	Spearman's rank correlation coefficient
	geographical distribution	GDP	Spearman's rank correlation coefficient
	geographical distribution	GDP per capita	Spearman's rank correlation coefficient
	geographical distribution	land mass	Spearman's rank correlation coefficient
	geographical distribution	geographical distance	Spearman's rank correlation coefficient
P4	Wikipedia coverage	GDP	Pearson correlation
	Data availability	Geographical location	Count & Mapping
	Data availability	Language Family	Count & Mapping
	Data availability	language class based on size and vitality	Count & Mapping
	Language model coverage	Geographical location	Count & Mapping
	Language model coverage	Language Family	Count & Mapping
	Language model coverage	Language class based on size and vitality	Count & Mapping
	Platform interface availability	Geographical location	Count & Mapping
	Platform interface availability	Language Family	Count & Mapping
	Platform interface availability	Language class based on size and vitality	Count & Mapping
	language class	Number of papers published	calculation of proportional share in sample
P8	Number of speakers	Number of papers published	Calculation of number of papers per million speakers
	Per capita	Number of papers published	Calculation of the highest paper count per capita

Table 9: Co-occurrence of two factors

ID	Resulting Value	Combination Method	Values Used
P1	Language occurrence entropy	Calculation of probability distribution of papers mentioning the same language Calculation of entropy Calculation of a class-wise mean reciprocal Rank which orders the languages based on their frequency of being mentioned in a conference	Number of papers mentioning the language per conference Publication year of papers taxonomy of languages based on available data
	Closeness of entities	Entity Embedding Analysis: Creation of word vectors based on input from papers, Prediction of the context, which is here author, language and conference	Authors, languages and conferences per paper Title and abstract of papers
P2	Degree to which the global demand is met by available LT	Calculation of sum(demand per language x utility) of LT areas	demand per language utility per language
P3	geographical representativeness of NLP datasets	entity recognition and linking creation of dataset-country maps Calculation of percentage of all entities associated with the single countries Calculation of number of countries not represented in the dataset	entities geographical association of entities countries in which the language is spoken
P5	digital language support	Mokken Scale Analysis: Scaling the coverage of the languages per DLS category	top tools per DLS category Languages covered by tools
P6	Diversity	Reuse of utility metrics from paper 2	
	Equity	Calculation of Gini-Coefficient	cumulative task performance per language
	Inclusion	Measures the benefit per unit increase in cost (cost = decrease in throughoutput and memory saved) Calculation of a average benefit-cost ratio for each language per task	Throughput Memory saved benefit
P7	Digital Language Equality	Technological factors: Each language resource, dataset or tool in the ELG Catalogue for a given language obtains a score which corresponds to the sum of the weights of its relevant features; per language all scores are summed up Contextual Factors: Weighted mean based on the size of the language communities in different countries, normalisation of values to 0-1, mean of all contextual factors defined as the overall contextual score for a respective language	Tools Services Datasets Language models Computational grammars Lexical and conceptual resources Annual GDP, GDP per capita; Perc. of the ICT sector in the GDP, ICT service exports in Balance of Payment Total no. of students in relevant area, Percentage of foreigners attaining tertiary education No. of enterprises in the field of I & C Scores extracted to represent the legal status of a language in different countries Number of articles in Wikipedia Innovation Index, Number of papers about the language Total number of speakers; Total number of social media users, Percentage of social media users Perc. of households with broadband
P8	Share of scripts	Calculation of proportional share of words in the specific scripts in the vocabulary of the model	Vocabulary per model scripts

Table 10: Approaches combining several factors