Proceedings of the TDLE Workshop @LREC2022, pages 13–26
Marseille, 20 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

Introducing the Digital Language Equality Metric: Contextual Factors

Annika Grützner-Zahn, Georg Rehm
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
{annika.gruetzner-zahn, georg.rehm}@dfki.de

Abstract

In our digital age, digital language equality is an important goal to enable participation in society for all citizens, independent of the language they speak. To assess the current state of play with regard to Europe's languages, we developed, in the project European Language Equality, a metric for digital language equality that consists of two parts, technological and contextual (i. e., non-technological) factors. We present a metric for calculating the contextual factors for over 80 European languages. For each language, a score is calculated that reflects the broader context or socio-economic ecosystem of a language, which has, for a given language, a direct impact for technology and resource development; it is important to note, though, that Language Technologies and Resources related aspects are reflected by the technological factors. To reduce the vast number of potential contextual factors to an adequate number, five different configurations were calculated and evaluated with a panel of experts. The best results were achieved by a configuration in which 12 manually curated factors were included. In the factor selection process, attention was paid to data quality, automatic updatability, inclusion of data from different domains, and a balance between different data types. The evaluation shows that this specific configuration is stable for the official EU languages; while for regional and minority languages, as well as national non-official EU languages, there is room for improvement.

Keywords: Language Technology, Digital Language Equality, Contextual Factors, Europe

1 Introduction

The rising influence of the internet on the daily life impacts the relevance of the automated understanding and production of language in the digital age since natural language is an important part of human-computer-interaction (HCI). From a technological perspective, Language Technologies (LT) can add the "ability to analyze, understand and generate information expressed in natural language" (Aldabe et al., 2021, p. 13) to digital systems. Especially many languages with smaller numbers of speakers are typically under-served in terms of resources and technologies, because of factors as missing economic interest, etc.. To analyse the situation of a language community in the digital sphere, it is necessary to develop a metric which is able to assess the current state of technological support, but that is also able to position the results in the broader socio-economic context of a language and its community. Hence, our suggested Digital Language Equality (DLE) metric consists of two broader groups of factors, *technological* and *contextual* factors.

Especially in multilingual societies, the importance and relevance of DLE is growing every day. In Europe, we are still far away from the ideal situation of DLE which would be "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." (Gaspari et al., 2021, p. 4). Back in 2012, the META-NET White Paper Series (Rehm and Uszkoreit, 2012) demonstrated a strong imbalance in terms of technology support for 31 European languages, even though at least the 24 official EU Member State languages have the same status and rights. Additionally, more than 60 regional and minority languages (RML) are protected via the European Charter for Regional or Minority Languages and the Charter of Fundamental rights of the EU (Article 21), which declare the prohibition of discrimination grounded on language (European Union, 1992; European Union, 2010).

The EU-funded project European Language Equality (ELE) addresses the challenge how to solve this existing imbalances. Its main goal is the preparation of a strategic research, innovation and implementation agenda and roadmap that specifies the necessary steps and instruments to achieve DLE in Europe by 2030 (Rehm and Way, 2022). The project covers a total of 89 languages: all 24 official EU languages, 11 official national languages without an official status in the EU and 54 regional and minority languages.[1]

For the preparation of the strategic agenda, the current state of each language needs to be determined as the starting point. In all previous attempts, such as the META-NET White Paper Series, the role of a language's context on the development (or lack thereof) of technologies for that language has been neglected. Accordingly, in this article we focus upon the contextual factors (CF).[2] We prepare different configurations of the metric based on simple classifications of the CFs to assess which factors can and should be included.

Section 2 provides the theoretical background about DLE in Europe and the measurement of the context of languages in the digital world. Section 3 describes the data collection, preparation and calculation of the metric. Section 4 presents the results and evaluation. Section 5 discusses the work and its limitations.

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

[2] A complementary paper, Gaspari et al. (2022), focuses on the technological factors.

2 Background and Related Work

2.1 Digital Language Equality in Europe

Digitisation brings people closer together and increases contact across national borders. For interaction across borders to function properly, smooth communication must be possible. However, communication has so far been dominated by a few languages with large communities of speakers or economic dominance. This excludes other language communities and can eventually lead to the digital extinction of a language. To avoid this scenario, smaller languages need to be supported, i. e., DLE must be defined as a societal, political and also scientific goal to enable languages to live and grow in the digital world.

Moreover, a multilingual society without proper translation has consequences. Negative impacts include not knowing certain information due to a lack of information access, no access to digital services in critical domains such as health and e-government, reduced and hindered participation in political processes and differences in cross-border shopping behavior (STOA, 2017; Burchardt et al., 2012; Bali et al., 2019). To avoid these effects, language barriers must be lowered or fully removed. With more than 80 languages in Europe, LTs are the only feasible option.

A recent European Parliament (EP) resolution acknowledges multilingualism to be a property of European diversity. Although it recommends setting up a "large-scale, long-term coordinated funding programme" (European Parliament, 2018) to decrease the differences between the technological support of Europe's languages, an EU policy to challenge language barriers does not exist yet (Aldabe et al., 2021). Additionally, language data, the foundation for the development of LTs, is not classified as "high value data" in the "Directive on open data and the re-use of public sector information", which implies that language data does not provide any benefit to society or economy, which is the main criterion for the classification (European Parliament and Council of the European Union, 2019). This creates an obstacle for LT development.

There are also differences in terms of research on different languages. English is better supported through LTs and is worked upon much more intensively than other languages in published work (Joshi et al., 2020; Blasi et al., 2021; Mager et al., 2018). In Europe, there has been more and more research on languages other than English in the last 10 to 15 years but the overall situation still cannot be considered one of equality.

Krauwer (2003) was one of the first calls for action towards the development of LRs/LTs for under-resourced languages. In the following years, different projects and initiatives established an important resource and technology basis for Europe's languages including, among others, Euromatrix (EU Publications Office, 2017a), FLaReNet (Soria et al., 2012), ITranslate4 (EU Publications Office, 2017b) and CLARIN (Hinrichs and Krauwer, 2014). Additionally, META-NET, an EU Network of Excellence forging the Multilingual Europe Technology Alliance, was established with a group of projects (T4ME, CESAR, METANET4U, METANORD) promoting and supporting the development of LTs for all European languages (Rehm and Uszkoreit, 2012; Rehm et al., 2014). The EU-funded project CRACKER (Cracking the language Barrier, 2015-2017) continued the work of META-NET, concentrating on additional strategy development and community building. The most recent EU projects in this line of actions are European Language Grid (ELG; Rehm et al., 2020a; Rehm, 2022) and European Language Equality (ELE; Rehm and Way, 2022). ELG and ELE collaborate closely, e. g., the DLE metric, developed in ELE, will be presented in a dedicated dashboard, which will be available in ELG.

In 2017, the report *Language equality in the digital age* was published (STOA, 2017), based on a study conducted by the Scientific Foresight Unit from the European Parliamentary Research Service. This report increased the awareness of the negative impacts of language barriers. LTs were proposed to be a possible answer, but, due to less funding and missing awareness, the danger of digital language extinction still threatens many European languages. Another problem identified was the lack of policies for LTs at national and European level (Rehm et al., 2020b). One year later, the *Language Equality in the digital age* resolution was adopted by the EP, which defines multilingualism as part of our cultural heritage and worthy of protection, as well as a challenge for an inclusive EU. It calls for the legal protection of the 60 European RMLs (European Parliament, 2018).

2.2 Measuring a Language's Context

Recently, research has begun to use data-driven approaches to establish relationships between the technical support of a language and non-technological factors, e. g., by clustering languages according to the number of available LRs and mentions in scientific publications. Joshi et al. (2020) show a correlation between the representation of a language at NLP conferences and the availability of language data. Mentions of each language in conferences were computed using Language Occurrence Entropy. Subsequently, a class-wise Mean Reciprocal Rank calculated the results per class in the conference proceedings.

Blasi et al. (2021) examine the performance of technologies for various languages as well as the correlation of technological and non-technological factors. Leaving technological performance aside, the authors analyse the correlation between the number of citations and the covered language diversity in a paper and between the economic size and number of published papers. A marginal effect was measured between the number of citations and number of languages covered by a paper, i. e., no correlation was detected. Significantly fewer prediction errors were found when the

Gross Domestic Product (GDP) was associated with the number of papers.

Moreover, data sets are also investigated regarding the correlation between geographical or economic factors and the origin of the data set calculating the predictive values for these factors (Faisal et al., 2021).

The AI Vibrancy Tool published with the AI Index report (Stanford University, 2021) computes a score that represents the "AI vibrancy" per country including TFs and non-technological factors. The factors covering research and development are based on numbers about publications, patents, AI conferences and available software. Economy is quantified via numbers about skills, hiring, investment and companies. Inclusion is represented through numbers about women in AI. The measured factors represent the context of AI software development (Zhang et al., 2021a). The calculation consists of the following steps: (1) data normalisation using a scalar; (2) calculation of the arithmetic average per country and indicator; (3) substitution of the values for each country into the formula[3]. Weights are applied to individual scores based on the respective indicator and the area of the indicator. Finally, for each factor a relative score between 0 and 100 is calculated (Stanford University, 2020).

In recent years, first approaches have been made to measure the technical support of languages. But due to the lack of data and the high complexity of the matter, a metric which includes all components is still missing. Section 3 shows that our DLE metric is based on a similar approach as the AI Index meaning it also results in a score through processing a number of factors and it is quantitative and solely data-driven.

3 Method

3.1 Data Collection

The preliminary definition of the DLE metric (Gaspari et al., 2021) included 72 potential contextual factors, clustered into 12 classes representing different aspects of the context of a language. Each of the factors had to be quantified with an indicator to be measurable, which depended on the presence and accessibility of data for a fitting indicator to represent the factor. First, different sources of pan-European data were collected. The selected ones included, among others, EUROSTAT[4], the European Language Monitor[5], Ethnologue[6] and various reports and articles. Second, the data was collected manually for each indicator.

Overall, 27 of the 72 initial factors were excluded due to missing data. This affected especially factors from the classes "research & development & innovation", "society" and "policy". Data about policies is mainly too broad and represents whether policies exist or not. The class "society" included factors about diversity being difficult to quantify. The problem of missing data in this area was already mentioned in the AI Index report (Zhang et al., 2021b). The factors excluded from the class "research & development & innovation" covered mainly figures about the LT research environment, while broader numbers about the research situation of the whole country were indeed available. Table 3 in the Appendix shows all factors from the preliminary definition (Gaspari et al., 2021), their class and the indicator they were quantified with. Overall, 46 factors[7] were quantified with at least one appropriate indicator, some with two indicators representing different perspectives like total numbers and numbers per capita.

The data was collected on 16 of Dec. 2021. Many sources provide their data as Excel sheets. Some data was published on websites. The data for 15 indicators had to be collected manually from reports or articles. We attempt to update the contextual factors on an annual basis. Based on the work presented in this paper, we assume that this process of updating the metric takes approximately one or two weeks.

3.2 Data Preparation

The collected data was very heterogeneous: it had different formats, was based on country or language level, included differing languages or countries and consisted of three data types. Data preparation took several steps, including the standardisation of the format of the numbers, harmonising the names of the languages (Hammarström et al., 2021) and merging the data from different tables. Some sources provided plain text from which a score had to be extracted. Features mentioned in the text were quantified with a score and this score was assigned to countries or language communities. If the text included more than one feature, the scores were added up. For a list of the indicators transformed from plain text and an explanation of the process see Table 4.

Because the metric is intended to process data on a language basis, data collected on the country level had to be converted to the language level. In total, the factors were quantified with three different types of data, total numbers, proportional numbers, and scores. Most total numbers were split proportionally, using the percentage of speakers of the language per country. The figures for the percentages were calculated through the population size and the number of speakers from Ethnologue.[8]

Due to some gaps and old records about RMLs, experts on minority languages from the ELE consortium were asked to fill the gaps or to provide better data. The figures for Alsatian, Faroese, Gallo, Icelandic, Macedonian and the Saami languages were corrected.

Percentages of languages often taught as a second language (English, German, French, Spanish) were only included if the language had an official status in the country. For example, the figures for English are based

[3] https://aiindex.stanford.edu/vibrancy/

[4] https://ec.europa.eu/eurostat

[5] http://www.efnil.org/projects/elm

[6] https://www.ethnologue.com

[7] The factor "political activity" was added.

[8] https://european-language-equality.eu/languages/

on the figures of the UK, Ireland and Malta. In other European countries, English does not have an official status, so they were not taken into account. If the language was an official national language in at least one country, only language communities with more than one percent were included to simplify the mapping. This calculation was performed for each language community in each of the European countries covered by the ELE project.
Total numbers per capita, proportional numbers, and scores were applied to the language communities without adjustment due to the complexity and additional time the adaption would have needed. A complex mapping would be desirable, as many language communities deviate from the average. Additionally, the mapping through the proportion of the speakers is problematic, too, because the sum of the speaker communities is not 100% if the country has many bi- or multilingual speakers. Hence, numbers from such countries were given several times. Another problem is the missing inclusion of the political reality regarding the promotion of a language. This refers to figures as to how many researchers work on the language, which were also transferred by a percentage mapping. In countries with a high number of speakers of a language, but less money or activity being spent on the promotion of the language, a direct mapping does not fit.
If a language was spoken in more than one country, total numbers were added up, while proportional numbers, scores and total numbers per capita were calculated through the average. At this point the different sizes of the language communities were slightly taken into account, hence, the data values of bigger language communities were weighted double for the calculation of the average. A complex inclusion of the size of the language community would result in more fine-grained figures and, therefore, probably in different scores.

3.3 Metric Calculation

The data per language community was converted into scores that indicate if a language has a context with the possibility to evolve or not. Without the political will, funding, innovation and economic interest in the region, the probability to achieve DLE is low. In order for the contextual values to be easy to compare and memorise, a score between 0 and 1 was assigned to the languages. Here, 0 represents a context with no potential for the development of LT, while 1 represents the best potential. To keep the metric as transparent as possible, it was decided to base the calculation on an average of the factors. Therefore, the intermediate goal was to calculate a score between 0 and 1 for each factor. The language with the lowest value for the respective factor will be depicted with a 0, the language with the highest value with a 1. The steps were as follows:

1. Calculation of range: highest value - lowest value;
2. $\frac{(value-minimum)*100}{range}$ = Percentage weighting of a language within the range;
3. The result is a relative value: to obtain a score between 0-1 the result is divided by 100;
4. Apply steps 1-3 for all languages and factors;
5. Calculate average of all factors per language;
6. Weighting of the scores with the three factors number of speakers, scores based on the language status and whether the language was an official language of the EU or not.

The three weighting factors were considered to be relevant for the context to develop LTs due to the influence of the number of speakers on the investment by large companies and the legal or EU status on the amount of funding. The weighting included two steps: 1) the calculation of the average of the overall scores, the scores for the number of speakers and the legal status and 2) the addition of 0.07 to the score for each official EU language. The second step was separated from the average calculation, because the indicator consisted of two values, 1 for being and 0 for not being an EU language. Average calculation would result in a too strong boost for the official EU languages. Hence, English had already a score of around 0.7 and 0.8 without the boost, smaller values for EU languages would have penalised English, which would not represent reality.
To create five different versions of the metric, the factors were classified based on the option to update the data automatically and the quality of the data (Table 3, indicators marked with * are automatically updateable and indicators marked with ** provides data with good quality). Data quality was chosen to avoid bias in the outcome of the metric. The possibility to update the data automatically was selected because it would simplify the implementation of the DLE metric in the form of an interactive dashboard in the ELG platform.
Based on these criteria, the following configurations of contextual factors were examined:

1. Factors with available data: 46 factors
2. Factors that can be updated automatically: 34 factors
3. Factors with good data quality: 26 factors
4. Factors that can be updated automatically and that have good data quality: 21 factors
5. Factors were manually curated using four criteria: automatically updatable, good data quality, not more than two factors per class, balance between data types: 12 factors (Table 1 shows the factors included in this configuration)

The fewer factors included in the metric, the more likely it is that an important influencing factor is omitted. However, the risk of distorting the metric with more data is reduced.

Table 1: Factors included in Configuration 5

Class	Factor
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

3.4 Heuristic Expert Evaluation

The results were validated through a heuristic expert evaluation, a method developed by the HCI community to conduct usability analyses. Experts were confronted with an interface and asked to give their opinion. One issue of the method is the lack of reproducibility, as differing opinions between experts produce different results. However, this allows for independent thoughts and maximises the likelihood of discovering aspects not noticed before (Nielsen and Molich, 1990). When three to four experts evaluate an interface together using this method, only 25-50% of errors are detected but with five independent experts between 55 and 90% of errors can be discovered (Georgsson et al., 2014).

We adapted the method for our purposes. The experts did not receive an interface but the results of the five configurations of the metric. The expert panel consisted of ELE consortium partners. The choice of the experts were based on their knowledge in the area of Language Technology, Computational Linguistics, Linguistics, Computer Science and others. Moreover, the experts represent different European countries and know the background of their countries and the languages spoken in the country well. We reached out to 37 (of the, in total, 52) ELE partners from 33 different organisations. The experts were asked to provide an intuitive assessment of the results regarding the languages they know, a feedback explaining how and why they would have expected the results to be and to indicate the most appropriate configuration.

4 Results

4.1 Most Adequate Configuration

The fifth configuration (Figure 1) was evaluated by the experts as being the one that reflects reality most adequately. The results of the other configurations are shown in the Appendix (Figures 2). Overall, the results develop steadily from the first configuration to the fifth in direction of higher scores for the official EU languages and lower scores for the regional and minority languages. From the second to the fifth configuration the results are similar but differ in the score ratios between the language groups (1) official EU languages, (2) national languages but not an official EU language and (3) regional and minority languages.

Diving deeper into the results of the fifth configuration (Figure 1), the calculated scores for the 89 languages with 12 curated factors range between 0.95 and 0. The distinction of 0.05 between the average of 0.14 and the median of 0.09 represents a left shift towards the higher scores. The first third is dominated by the official EU languages (turquoise) ranging between a score of 0.17 and 0.95, while the RMLs (orange) are presented as a long tail with low scores between 0.1 and 0. The official national languages which are not recognised as official EU languages (pink) are between the other two language groups having scores from 0.18 to 0.08. The proximity of English, German and French and the relatively low score for Spanish are caused by the inclusion of only European countries in the data.

Generally, the results exhibit a Northwest to Southeast divide. Usually, the languages spoken in the Northwest of a language group have better scores than the languages spoken in the Southeast of Europe. This tendency materialises especially in the regional and minority languages and less in the official EU languages.

4.2 Heuristic Expert Evaluation

From the 37 contacted partners, 18 provided an assessment of the results. The feedback consisted of overall ratings of the five configurations (Section 4.1) as well as detailed comments regarding individual languages the experts have expertise in. As a consequence, most answers related to official EU languages. RMLs for which feedback was received are spoken in the UK, Spain, Italy and the Nordic countries. We received feedback on 56 of the 89 languages.

In general, using all factors was evaluated as risky due to the possible distortion of results caused by data with bad quality. The results of configuration 1 were indeed considered as being counterintuitive, with high scores for languages as Emilian, Gallo and Franco-Provencial which seemed to be motivated by distorted data. The second configuration was similarly criticised, except for positive comments on the automatic nature of the metric. The results are less distorted but evaluated as worse compared to configurations 3-5. The results of the third and fourth configuration are similar. Focusing on quality data improves the results significantly, but fewer factors eventually imply that relevant important factors for the context may be missing. However, although the factors were reduced the scores remain similar. The fifth configuration was assessed positively regarding the transparency of fewer factors and the possibility to balance the factor classes.

The evaluation of individual languages and their scores showed an improvement from the first configuration with the worst results to the fifth configuration with the best results. Table 2 lists the evaluated languages in

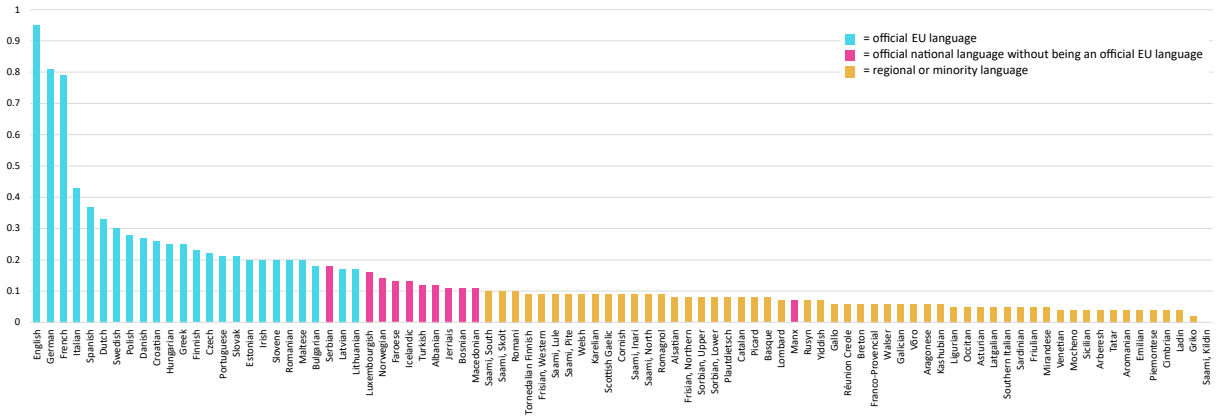


Figure 1: Results of Configuration 5 (12 manually curated factors)

configuration 5 and their assessment by the experts.

Overall, the results of the fifth configuration were assessed to represent the context of the language communities in the most adequate way, while there is still room for improvement for a few languages.

Various experts provided suggestions to improve the results, e. g., with regard to the data and data sources. First, it was recommended to collect data in national and regional sources. Additionally, it was pointed out that the context of languages spoken outside of Europe is excluded and therefore important and relevant numbers are missing. Other suggestions refer to missing factors, particularly relevant for RMLs is the inclusion of the vitality status of a language. Another idea was to replace the official EU status as a weighting factor with the respective country's membership in the European Economic Area (EEA), since countries within this alliance have access to European research funds and networks. Moreover, competition between official national languages, as with Irish and English in Ireland, was suggested to be an important factor.

There were also suggestions regarding the presentation of the results. Language communities with complex political backgrounds are most likely to be biased by a rather simplistic calculation based on country data and should be highlighted and presented with the limitations of data-based approaches for such cases. It was also suggested that languages that do not have a writing system are special cases for the development of LTs.

A few experts had more global comments regarding the methodology, which they felt is unable to capture the complex contexts of certain language communities such as Maltese, Irish and the Celtic languages, which scored better than expected. The prosperity of the UK boosts its relevant RMLs with the country-specific data, while in reality these RML communities are strongly dominated by English. The same applies to Ireland and the Irish language. Another point of criticism was the inclusion of data not applied per capita. As a result, small language communities, despite relatively good support, cannot achieve a high score. The size of the language community has an impact on the economic interest, investment, number of researchers, etc. for the language, but for some smaller language communities that have already invested a lot in their language and infrastructure, the score seems too low.

5 Discussion

The DLE metric for the contextual factors has some limitations (see Section 3). First, expanding the data set to include regional or national sources would result in (i) a higher number of factors, (ii) improved data quality, as gaps in individual indicators may be filled, (iii) quantification of more factors with more than one indicator, to reflect different perspectives and (vi) a more complex mapping to language communities based on regional data resulting in a significant impact on RML. Moreover, including the factors suggested by the experts, such as membership in the EEA or language vitality status, could help improve the results.

Second, the data cleaning procedure can be improved. For the calculation of the Innovation Scoreboard (Bielinska-Dusza and Hamerska, 2021), outliers with values outside twice the standard deviation were replaced by the respective maxima or minima of the data series. Data gaps could be filled using data from previous years and skewed data could be corrected using a square root transformation. These steps would most likely affect the results of configurations 1 and 2, since they use the data with poorer quality.

The mapping of country-specific data to language-specific data can be improved, e. g., Bromham et al. (2021) show how a possible regional mapping of data using the World GeoDatasets[9] could be realised. For large countries with bigger regional or urban-rural divides, a regional mapping would represent reality more accurately. In particular, the missing mapping of proportional data, scores and total numbers per capita has a major impact on the results. Here, regional data could help to calculate the average deviation of individual regions or language communities from other proportional data and to transfer this deviation to proportional data

[9] http://worldgeodatasets.com

Suitable	Too high	Too low	Contrary Opinion
English	Irish	Norwegian	French
Dutch	Italian	Spanish	German
Danish	Swedish	Portuguese	Saami, Northern
Polish	Hungarian	Czech	Latvian
Greek	Croatian	Romanian	
Finnish	Maltese	Bulgarian	
Estonian	Faroese	Icelandic	
Slovene	Scottish Gaelic	Emilian	
Slovak	Cornish	Sicilian	
Lithuanian	Manx		
Serbian	Saami, Southern		
Basque	Saami, Pite		
Catalan	Saami, Lule		
Galician	Saami, Skolt		
Asturian	Saami, Inari		
Aragonese	Sardinian		
Welsh	Romagnol		
Griko			
Lombard			
Ligurian			
Venetian			
Southern Italian			
Friulian			
Piemontese			
Ladin			
25	17	9	4

Table 2: Assessment of the individual languages in configuration 5 by the panel of experts

only found on national level, and similarly for the total figures per capita. Another improvement would be to calculate the data merging from the individual language communities in different countries depending on the size of the language community. Currently, the values of larger language communities were double-weighted when determining the average of proportional data, numbers per capita or scores. This simplification could be mitigated by including the total number of speakers per language community in each country. Sustainability was mentioned several times. Romaine (2017, p. 49) stressed the importance of an "on-going monitoring of individual communities" for a reliable evaluation of the situation regarding language diversity which was considered in this approach as an important aspect and taken into account with the inclusion of the criterion automatic updateability of the factors. One problem for the future is the relative calculation from the values to each other. Thus, the scores of *all* languages may change if new values are added, even if the situation of the language community itself has not changed. To mitigate this, a temporal dimension could be integrated (Bielinska-Dusza and Hamerska, 2021). The lowest and highest value of the range for the calculation represent the lowest or the highest value from the last years, which reduces fluctuations.

Another approach would be to measure the prediction accuracy of the CFs with regard to the TFs after some time. In this way, each single factor could be evaluated and unrecognized distortion in the results could be examined and ruled out in the future.

The results show a need for an improvement regarding the context for LT development for all languages except English, French and German. Despite the lack of data about non-European countries with English as the official national language, English achieves the best results in every configuration. Thus, the dominance of English in business and science is reflected in the data. The good results for French and German are grounded in the size of the countries and their economies. Spanish reaches only half the score, even though it has many more speakers. Some experts criticise this result since the context of Spanish for LT development should score higher. As shown by the META-NET White Papers (Rehm and Uszkoreit, 2012), LT support for Spanish is similar to German and French. Since the CFs are supposed to show the achievability of DLE and thus give a 'prediction' for LT development, the results do not fit.

In the META-NET White Paper comparison of the technical support of Europe's languages, the languages that were assessed as having a better technical support in 2012 also perform better in the calculation of the CFs. Always reaching the highest contextual scores, English, Dutch, French, German, Spanish and Italian achieved "moderate support" in at least three of the four LT areas (Rehm and Uszkoreit, 2012). The next set of languages according to the results of the CFs, i. e., Polish, Czech, Swedish, Hungarian and Finnish,

also achieved "moderate support" in at least one area in 2012. The fact that these languages achieved better results in 2012 indicates that their context has probably been better ten years ago than for the remaining languages. Greek, Croatian and Danish stand out because these three languages did not reach the "moderate support" level in any of the four groups in 2012. However, since the score for Croatian is considered too high by the experts for all configurations, it can be assumed that the score is distorted by the data. The context for Greek and Danish seems to have improved.

Blasi et al. (2021) and Joshi et al. (2020) highlight the marginal representation in research of languages with a small language community and a low economic weight. The results based on an academic context are not deviating from results based on the entire context as presented in the present paper (Joshi et al., 2020). Additionally, Blasi et al. (2021) point out the more complex the technical task, the worse the technical support languages with a small number of speakers have, i. e., the size of the language community seems to have an influence on the technical support. Faisal et al. (2021) predict the correlation between data sets and the country of origin with three factors: GDP, size of the language community and geographic proximity. Most of the data sets came from economically prosperous countries, thus the best predictive value was the GDP. Additionally, Blasi et al. (2021) show that the GDP has a better predictive power regarding the publication of papers than the number of speakers of a language. According to these results, the GDP has a stronger influence in academia than the size of the language community. However, if language communities have both, a low GDP and few speakers, special effort and support are needed to ensure technical support.

According to the Northwest to Southeast divide identified, it is the context of language communities in the East and South of Europe that needs to be strengthened to achieve DLE. In the META-NET White Paper Series, only three languages spoken in Eastern Europe achieved "moderate support" once in the four areas. In comparison, the technical support of nine languages spoken in the West was rated as "moderate" at least once. Since no other related studies exist, these results can only be discussed in a broader context. For example, Bargaoanu et al. (2019) identified an East-West difference in Europe using data on economic and social development patterns. Although fewer factors were examined, the same pattern emerges. The difference between Northwest and Southeast needs to be reduced to enable all language communities to participate in the digital society. The results are particularly poor for small language communities. In order for the EU to be a truly equal association of countries and language communities, the differences must be evened out. Otherwise, the impact of language barriers (Section 2.1) will remain and even reinforce inequalities.

The results of the CFs along with the technological scores form the Digital Language Equality metric. Both scores will be presented in an interactive, web-based dashboard and will provide information about the current state of LT support based on the TFs and about the situation of the language communities regarding the further development. Together, TF and CF scores/results can be used as the basis for strategic recommendations regarding the future development of languages in the digital world. A language that is poorly supported technologically and has a bad contextual score is unlikely to exhibit significant improvement regarding LT support without changing its context. A language lacking LT support but with a better situational context could indeed take the next steps towards DLE in the coming years. Currently well-supported languages will continue to do well if their good situational context stays intact, while languages with a good technological score and a rather low context, are likely to stagnate technologically.

6 Conclusion

We present a first approach for the calculation of a score, which is meant to reflect the context of a language with respect to the development of LTs. The DLE metric consists of technological factors representing the current state of technical support and contextual factors describing the situation for LT development and achievability of DLE, especially with regard to the languages covered by ELE. The scores can also be used to create initial predictions about the further LT development if the context does not change.

Our initial methodological approach exhibits room for improvement. This applies in particular to the data collection and preparation. The mapping of data from the country to the language level can be improved, reducing inherent inaccuracies affecting data from language communities with few speakers. Another approach could be the calculation of predictive values for individual CFs based on TF scores. This would allow each individual factor to be tested for its predictive power regarding LT development.

The results of the five tested configurations show a clear pattern once they are reduced by the factors that distort the results due poor data quality. There exists a greater difference between the scores of the official EU languages and RMLs, as well as a gradient from Northwest to Southeast within the groups.

The heuristic expert evaluation has shown that the results of the fifth configuration correspond most closely to reality. The scores of some languages, especially those in a more complicated political environment, do not yet adequately represent their language community's context. These results can be improved using the suggestions presented. The result of this initial approach provides a first starting point from which further development regarding aspects as clarity and reproducibility can be pursued.

Acknowledgments

The work presented in this article was co-financed by the European Union under grant agreement no. LC-01641480 – 101018166.

7 Bibliographical References

Aldabe, I., Rehm, G., Rigau, G., and Way, A. (2021). D3.1 Report on existing strategic documents and projects in LT/AI. https://european-language-equality.eu/wp-content/uploads/2021/12/ELE___Deliverable_D3_1__revised_.pdf.

Bali, K., Choudhury, M., Sitaram, S., and Seshadri, V. (2019). ELLORA: Enabling Low Resource Languages with Technology. In *Proceedings of the 1st International Conference on Language Technologies for All*, pages 160–163, Paris, France. European Language Resources Association.

Bargaoanu, A., Buturoiu, R., and Durach, F. (2019). The East-West Divide in the European Union: A Development Divide Reframed as a Political One. In Paul Dobrescu, editor, *Development in Turbulent Times: The Many Faces of Inequality Within Europe*, pages 105–118, Cham. Springer International Publishing.

Bielinska-Dusza, E. and Hamerska, M. (2021). Methodology for Calculating the European Innovation Scoreboard - Proposition for Modification. *Sustainability*, 13(4).

Blasi, D., Anastasopoulos, A., and Neubig, G. (2021). Systematic Inequalities in Language Technology Performance across the World's Languages. https://arxiv.org/abs/2110.06733.

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., and Hua, X. (2021). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6:163–173.

Burchardt, A., Egg, M., Eichler, K., Krenn, B., Kreutel, J., Leßmöllmann, A., Rehm, G., Stede, M., Uszkoreit, H., and Volk, M. (2012). *Die Deutsche Sprache im digitalen Zeitalter – The German Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Springer, Heidelberg, New York, Dordrecht, London.

EU Publications Office. (2017a). EuroMatrix: Statistical and Hybrid Machine Translation Between All European Languages. https://cordis.europa.eu/project/id/034291. Last accessed: 07.02.2022.

EU Publications Office. (2017b). Internet Translators for all European Languages. https://cordis.europa.eu/project/id/250405. Last accessed: 07.02.2022.

European Parliament and Council of the European Union. (2019). Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32019L1024. Last accessed: 04.02.2022.

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. Last accessed: 02.02.2022.

European Union. (1992). European Charter for Regional or Minority Languages. *Council Of Europe European Treaty Series*, 148.

European Union. (2010). Charter of fundamental rights of the european union. *Official Journal of the European Union C83*, 53.

Faisal, F., Wang, Y., and Anastasopoulos, A. (2021). Dataset Geography: Mapping Language Data to Language Users. *Computing Research Repository (CoRR)*, abs/2112.03497. Last accessed: 10.02.2021.

Gaspari, F., Way, A., Dunne, J., Rehm, G., Piperidis, S., and Giagkou, M. (2021). D1.1 Digital Language Equality (preliminary definition). https://european-language-equality.eu/wp-content/uploads/2021/05/ELE_Deliverable_D1_1.pdf.

Gaspari, F., Gallagher, O., Rehm, G., Giagkou, M., Piperidis, S., Dunne, J., and Way, A. (2022). Introducing the Digital Language Equality Metric: Technological Factors. In Itziar Aldabe, et al., editors, *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*, Marseille, France. Accepted for publication. 20 June 2022.

Georgsson, M., Weir, C. R., and Staggers, N. (2014). Revisiting Heuristic Evaluation Methods to Improve the Reliability of Findings. *Studies in health technology and informatics*, 205:930–934.

Hammarström, H., Forkel, R., Haspelmath, M., and Bank, S. (2021). *Glottolog 4.5*. Max Planck Institute for Evolutionary Anthropology, Leipzig. Last accessed: 09.02.2022.

Hinrichs, E. and Krauwer, S. (2014). The CLARIN Research Infrastructure: Resources and Tools for eHumanities Scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1525–1531, Reykjavik, Iceland. European Language Resources Association.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (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.

Krauwer, S. (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.

Mager, M., Gutierrez-Vasques, X., Sierra, G., and Meza-Ruiz, I. (2018). Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Nielsen, J. and Molich, R., (1990). *Heuristic Evaluation of User Interfaces*, page 249–256. CHI '90. Association for Computing Machinery, New York, USA.

Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., Branco, A., Budin, G., Calzolari, N., Daelemans, W., Garabík, R., Grobelnik, M., García-Mateo, C., Hajič, J., Hernáez, I., Judge, J., Koeva, S., Krek, S., Krstev, C., and McNaught, J. (2014). The Strategic Impact of META-NET on the Regional, National and International Level. In *Language Resources and Evaluation*, volume 50.

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., Hajič, 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., and Klejch, O. (2020a). European Language Grid: An Overview. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3366–3380, Marseille, France. European Language Resources Association.

Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajic, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Pérez, J. M. G., Meertens, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Bars, L. L., Anastasiou, D., Auksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., Smedt, K. D., Garabík, R., Gavriilidou, M., Gromann, D., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Odijk, J., Ogrodniczuk, M., Rögnvaldsson, E., Rosner, M., Pedersen, B., Skadina, I., Tadić, M., Tufiș, D., Váradi, T., Vider, K., Way, A., and Yvon, F. (2020b). The European Language Technology Landscape in 2020: Language-Centric and Human-Centric AI for Cross-Cultural Communication in Multilingual Europe. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3315–3325, Marseille, France, 5. European Language Resources Association (ELRA).

Rehm, G. and Uszkoreit, H., editor. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc. Springer.

Rehm, G. and Way, A., editor. (2022). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer. Forthcoming.

Rehm, G., editor. (2022). *European Language Grid: A Language Technology Platform for Multilingual Europe*. Cognitive Technologies. Springer. Forthcoming.

Romaine, S. (2017). Language Endangerment and Language Death. In *The Routledge Handbook of Ecolinguistics*, pages 40–55. Routledge.

Soria, C., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., and Calzolari, N. (2012). The FLaReNet Strategic Language Resource Agenda. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1379–1386, Istanbul, Turkey. European Language Resources Association.

Stanford University. (2020). Global AI Vibrancy Tool. https://aiindex.stanford.edu/vibrancy/. Last accessed: 05.02.2022.

Stanford University. (2021). About: Developing a deeper understanding of a complex field. https://aiindex.stanford.edu/about/. Last accessed: 05.02.2022.

STOA. (2017). Language equality in the digital age – Towards a Human Language Project. http://www.europarl.europa.eu/stoa/. Last accessed: 13.01.2022.

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., and Perrault, R. (2021a). The AI Index 2021 Annual Report. https://aiindex.stanford.edu/report/. Last accessed: 05.02.2022.

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., and Perrault, R. (2021b). The AI Index 2021 Annual Report. https://arxiv.org/abs/2103.06312.

Appendix

Table 3: Initially proposed contextual factors (Gaspari et al., 2021)

Class	Factor	Indicator
Economy	Size of the economy	Annual GDP
		GDP per capita* **
	Size of the LT/NLP market	LT market in million Euro
	Size of the language service, translating or interpreting market	Number of organizations from the industry in the ELG catalogue* **
	Size of the IT/ICT sector	Perc. of the ICT sector in the GDP* **
		ICT service exports in Balance of Payment* **
	Investment instruments into AI/ LT	GDE on R&D in relevant areas*
	Regional/ national LT market	No indicator found
	Average socio-economic status	Annual net earnings, 1.0 FTE worker* **
		Life expectancy at age 60**
Education	Higher Education Institutions operating in the language	No indicator found
	Higher education in the language	No indicator found
	Academic positions in relevant areas	Head count of R&D personnel
	Academic programmes in relevant areas	No indicator found
	Literacy Level	Literacy rate*
	Students in language/LT/NLP curricula	Total no. of students in relevant areas* **
	Equity in education	Proportional tertiary educ. attainment* **
	Inclusion in education	Percentage of foreigners attaining tertiary education* **
Funding	Funding available for LT research projects	No. of projects funded in relevant areas*
		Score from the National funding programs
	Venture capital available	Venture capital amounts in Euro
	Public funding for interoperable platforms	Number of platforms**
Industry	Companies developing LTs	No. of enterprises in the field of I & C* **
	Start-ups per year	Percentage of "Enterprise births"**
	Start-ups in LT/ AI	Number of AI start ups* **
Law	Copyright legislation and regulations	No indicator found
	Legal status and legal protection	Scores out of the legal status* **
Media	Subtitled or dubbed visual media outcomes	Scores out of language transfer practices*
		Scores out of answers about broadcast practices
	Transcribed podcasts	Number of entries in the cba*
Online	Digital libraries	Percentage of contribution to Europeana
	Impact of language barriers on e-commerce	Percentage of population buying cross-border**
	Digital literacy	No indicator found
	Wikipedia pages	Number of articles in Wikipedia* **
	Websites exclusively in the language	No indicator found
	Websites in the language (not exclusively)	Perc. of websites in the languages* **
	Web pages	No indicator
	Ranking of websites delivering content	12 selected websites supporting the languages
	Labels and lemmas in knowledge bases	Number of lexemes in Wikipedia* **
	Language support gaps	Language matrix of supported features*
	Impact on E-commerce websites	T-Index*

Continued on next page

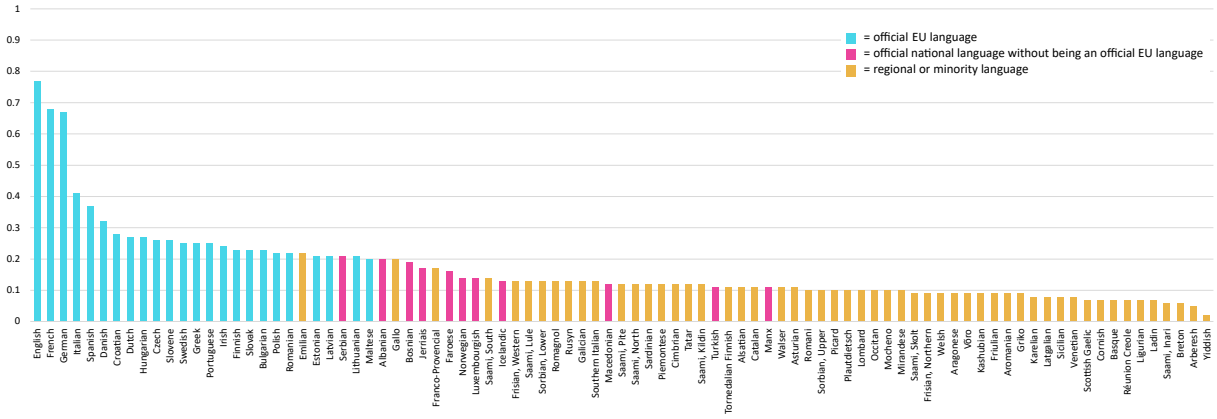
Table 3 – *Continued from previous page*

Class	Factor	Indicator
Policy	Presence of strategic plans, agendas, etc.	Scores out of a list of the published national AI strategies
		Scores from questionnaire about strategies
	Promotion of the LR ecosystem	No indicator found
	Consideration of bodies for the LR citation	No indicator found
	Promotion of cooperation	No indicator found
	Public and community support for resource production best practices	No indicator found
	Policies regarding BLARKs	No indicator found
	Political activity	Scores out of the list of documents
Public administration	Languages of public institutions	No. of constitutions written in the language
	Available public services in the language	Percentage of a maximum score about digital public services**
		Score for digital public services**
Research & Development & Innovation	Innovation capacity	Innovation Index* **
	Research groups in LT	Number of research organizations
	Research groups/ companies predominantly working on the respective language	No indicator found
	Research staff involved in LT	No indicator found
	Suitably qualified Research staff in LT	No indicator found
	Capacity for talent retention in LT	No indicator found
	State of play of NLP/AI	No indicator found
	Scientists working in LT/ on the language	Number of researchers in relevant areas*
	Researchers whose work benefits from LRs and LTs	No indicator found
	Overall research support staff	Head count of research support staff* **
	Scientific associations or general scientific and technology ecosystem	No indicator found
	Papers about LT and or the language	Number of papers about LT**
		Number of papers about the language* **
Society	Importance of the language	No indicator found
	Fully proficient (literate) speakers	Number of L1 speakers*
	Digital Skills	Perc. of individuals with basic digital skills* **
	Size of language community	Total number of speakers* **
	Population not speaking the official language(s)	No indicator found
	Official or recognized languages	Total no. of languages with official status*
		Number of bordering languages
	Community languages	Number of community languages*
	Time resources of the language community	No indicator found
	Society stakeholders for the language	No indicator found
	Speakers' attitudes towards the language	Total number of participants wanting to acquire the language
	Involvement of indigenous peoples	No indicator found
	Sensitivity to barriers	No indicator found
	Usage of Social Media or networks	Total number of social media users* **
		Percentage of social media users* **
Technology	Open-source technologies of LTs	No indicator found
	Access to computer, smartphone etc.	Perc. of households with a computer* **
	Digital connectivity and Internet access	Perc. of households with broadband* **

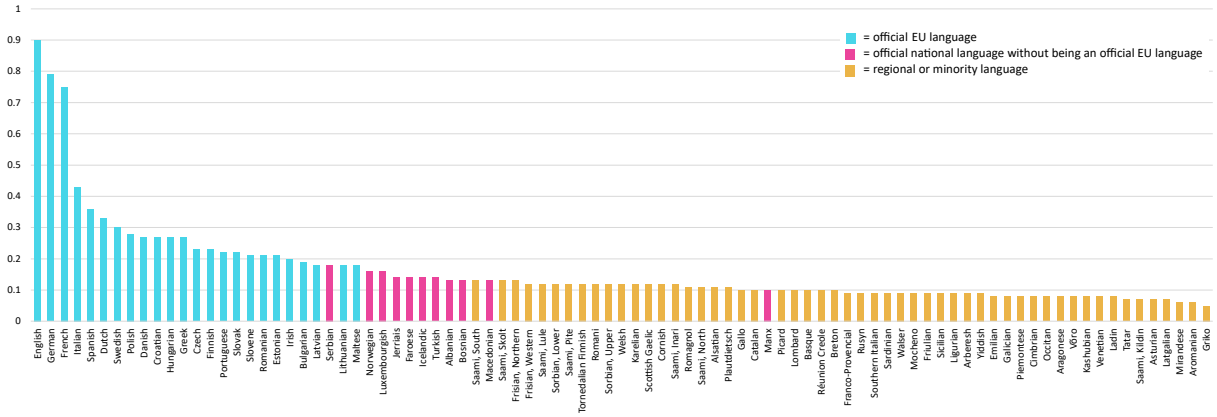
*Indicator marked * is automatically updateable – Indicator marked ** provides data with good quality*

Table 4: Conversion from plain text to scores

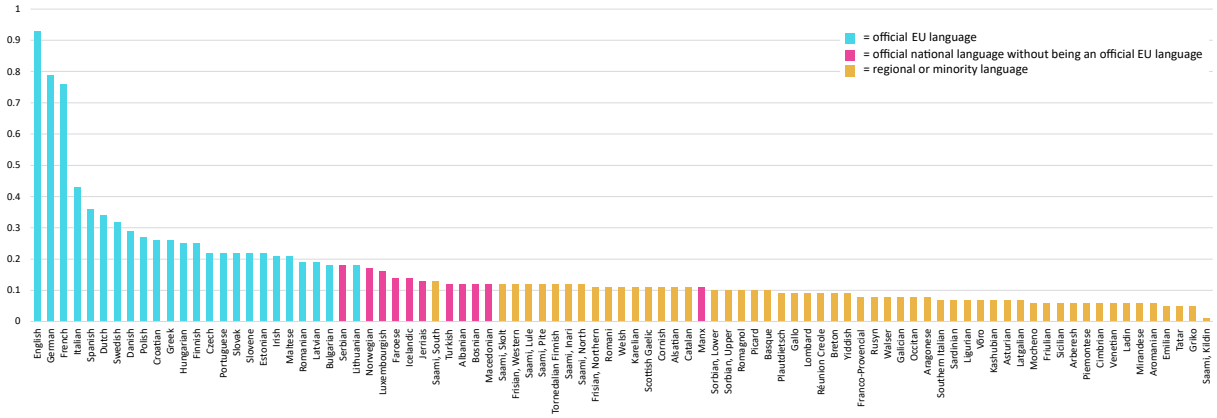
Factor	Merging of scores	Conversion from Text to Scores
Public funding available for LTs	Adding up of the scores for each country	1 for regional funding 1 for national funding 1 for intranational funding 1 each for ESIF, EUREKA, EUROSTAT
Legal status and legal protection	Adding up of the scores per language	10 for statutory national language 10 for de facto national working language 2 for statutory provincial language 2 for statutory provincial working language 1 for recognized language
Publicly available media outcomes	Sum of two scores: one for language transfer practices for films screened, one for tv broadcasts	2 for dub 1.5 for voice over 1.5 for sub and dub 1 for sub
	Sum of the scores + division through number of answers	Broadcast in original language: 5 for mostly/ always, 2.5 for sometimes . . . with dubbing: 4 for mostly/ always, 2 for sometimes . . . in original language with voice-over: 3 for mostly/ always, 1.5 for sometimes . . . with subtitles: 1 for mostly/ always, 0.5 for sometimes Dual-channel audio: 2 for mostly/ always, 1 for somet.
Presence of local, regional or national strategic plans	One of the score per country	1 for no plan/ strategy 2 for a plan without mentioning LT 3 for a plan mentioning LT 4 for a plan mentioning LT, minority, regional languages
Political activity	Adding up of the scores per country	1 score for each document (mentioning LT) 2 for each document exclusively about LT 1 for a document covering a specific language 2 for each document published 2020/2021 1 for each document published 2019/2018



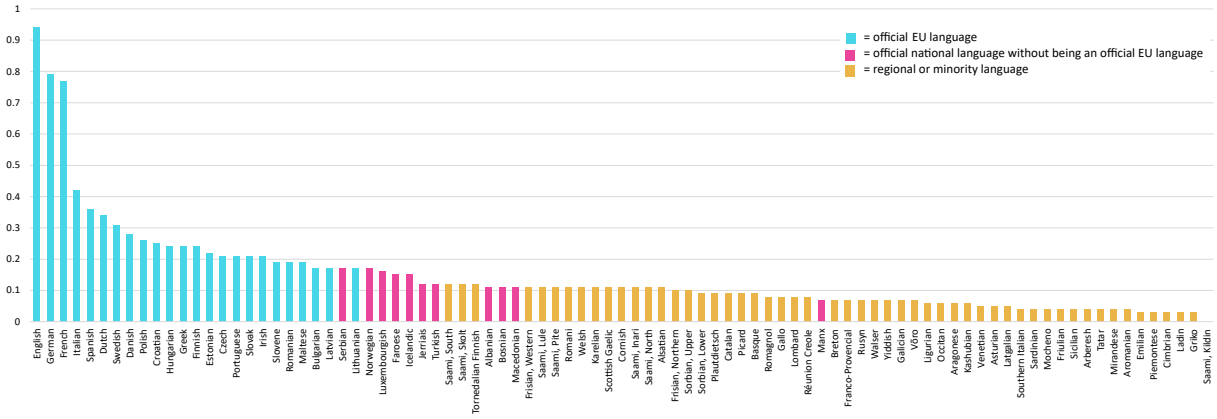
Results of Configuration 1 (46 factors with available data)



Results of Configuration 2 (34 factors that can be updated automatically)



Results of Configuration 3 (26 factors with good data quality)



Results of Configuration 4 (21 factors that can be updated automatically and that have good data quality)

Figure 2: Results of Configurations 1 to 4