Proceedings of the TDLE Workshop @LREC2022, pages 1–12
Marseille, 20 June 2022

Introducing the Digital Language Equality Metric: Technological Factors

Federico Gaspari[1], Owen Gallagher[1], Georg Rehm[2], Maria Giagkou[3], Stelios Piperidis[3], Jane Dunne[1], Andy Way[1]
[1]ADAPT Centre, School of Computing, Dublin City University, Ireland
[2]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Germany
[3]Institute for Language and Speech Processing, Research Centre "Athena", Greece
{firstname.lastname}@adaptcentre.ie, {firstname.lastname}@dfki.de, {mgiagkou, spip}@athenarc.gr

Abstract

This paper introduces the concept of Digital Language Equality (DLE) developed by the EU-funded European Language Equality (ELE) project, and describes the associated DLE Metric with a focus on its technological factors (TFs), which are complemented by situational contextual factors. This work aims at objectively describing the level of technological support of all European languages and lays the foundation to implement a large-scale EU-wide programme to ensure that these languages can continue to exist and prosper in the digital age, to serve the present and future needs of their speakers. The paper situates this ongoing work with a strong European focus in the broader context of related efforts, and explains how the DLE Metric can help track the progress towards DLE for all languages of Europe, focusing in particular on the role played by the TFs. These are derived from the European Language Grid (ELG) Catalogue, that provides the empirical basis to measure the level of digital readiness of all European languages. The DLE Metric scores can be consulted through an online interactive dashboard to show the level of technological support of each European language and track the overall progress toward DLE.

Keywords: Digital Language Equality, Technological Factors, Language Resources, Tools, Technologies, Europe

1 Introduction and Background

1.1 Motivation and Objectives

In a plenary meeting on 11 September 2018, the European Parliament adopted by an overwhelming majority a joint ITRE/CULT report, *Language equality in the digital age*, with a resolution that included over 40 recommendations. These concerned *inter alia* the enhancement of the institutional framework for Language Technology (LT) policies at EU level, as well as of EU research and education policies to improve the future of LTs in Europe, so that all stakeholders could benefit from them (European Parliament, 2018).

In an effort to address these recommendations, the European Language Equality (ELE) project[1] (Rehm et al., 2022; Rehm and Way, 2022) with its 52-member consortium is engaged in responding to the call to establish a much-needed large-scale, long-term coordinated funding programme for research, development and innovation in the field of LTs, at European, national and regional levels, designed to meet Europe's needs and demands. By addressing some of the key recommendations issued by the European Parliament, ELE is laying the foundation to draw up an evidence-based Strategic Research, Innovation and Implementation Agenda (SRIA) and Roadmap with strong support from the wider community, as a basis to launch a large-scale programme to achieve full Digital Language Equality (DLE) in Europe by 2030.

The ELE consortium is ideally positioned to pursue this ambitious objective, in that its members include a combination of research and academic organisations, networks, associations and initiatives as well as companies from all over Europe. In addition to all official European languages, the partners' combined expertise covers a very wide range of regional and minority languages, either through consortium partners or through several umbrella organisations.

1.2 Current Situation and Related Work

While the ongoing work conducted by ELE is focused on the languages of Europe, it is situated in a broader context of recent similar efforts with a wider remit. Joshi et al. (2020) investigate the relation between the languages of the world and the resources available for them as well as their coverage in Natural Language Processing (NLP) conferences, providing evidence for the severe disparity that exists across languages in terms of technological support and attention paid by academic, scientific and corporate circles.

Blasi et al. (2021) argue that the substantial progress brought about by the generally improved performance of NLP methods "has been restricted to a minuscule subset of the world's 6,500 languages", and present a framework for gauging the global utility of LTs in relation to demand, based on the analysis of a sample of over 60,000 papers from all major international NLP conferences. This study also shows convincing evidence for the striking inequality in the development of LTs across the world's languages. While this severe imbalance is partly in favour of a few, mostly European, languages, on the whole most European languages are at a disadvantage. Acknowledging that LTs are generally becoming increasingly ubiquitous, Faisal et al. (2021) look into the efforts to expand the language di-

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

versity and coverage of NLP applications. Since a key factor determining the quality of present-day NLP systems is data availability, they study the geographical representativeness of language datasets, to assess the extent to which they match the needs of the members of the respective language communities, with a thorough analysis of the striking inequalities.

Bromham et al. (2021) examine the effects of a wide range of demographic and socio-economic aspects on the use and status of the languages of the world, and reach the conclusion that language diversity is under threat across the globe, including in industrialised and economically advanced regions. In particular, this study found that half of the languages under investigation face serious risks of extinction, potentially within a generation, if not imminently. This is certainly a very sombre situation to face up to, which calls for a large-scale mobilisation of all possible efforts by all interested parties to avoid such a daunting prospect, especially for the languages addressed by ELE.[2] It should be emphasised that ELE covers not only the official languages of the European Union or national languages, but also regional and minority languages, and in fact these receive special attention insofar as they are among the least resourced and those with more limited technological support, which puts their communities at a serious disadvantage in the digital age.

1.3 Structure of the Paper

The rest of this article is organised as follows. Section 2 explains the principles behind the Digital Language Equality concept adopted in ELE and the rationale for the DLE Metric with an emphasis on the Technological Factors (TFs). Section 3 zooms in on the TFs, which are complemented by the Contextual Factors (CFs), outlining their main components and discussing the role of the European Language Grid (ELG) as its empirical evidence. The weights assigned to the feature values of the TFs are described, reporting on the main findings of the experiments that were conducted to refine the first implementation of the DLE Metric. The discussion emphasises the flexibility of the DLE Metric, that can be adapted in the future to accommodate subsequent developments and novelties in the community that it may not be possible to anticipate at present. We present our initial results regarding the current level of technological support and digital readiness of Europe's languages based on the TFs of the DLE Metric (the Technological DLE score), computed using a weighting scheme. We also briefly review some of the main open issues and challenges that remain to be addressed. Finally, Section 4 draws some conclusions, pointing out the value and potential of the DLE Metric to benefit the wider LT community and, ultimately, the European citizens on the whole by supporting their future aspirations in the digital age.

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

2 Digital Language Equality Metric

2.1 Guiding Principles

This paper introduces the notion of Digital Language Equality (DLE) developed in the project ELE to pursue its ambitious objectives, and presents the associated metric, focusing in particular on the Technological Factors (TFs). The DLE definition is intended to serve the needs of the languages in scope of ELE and the expectations of the relevant language communities in the future. It should be noted that language "equality" does not mean "sameness" on all counts, regardless of the respective environments; we recognise the different historical developments and current situations of the very diverse languages targeted in and by the project, along with their specific features, different needs and realities of their communities, e. g., in terms of number of speakers, ranges of use, etc., which inevitably vary significantly. It would be naive and unrealistic in practice to ignore these facts, and to set out to erase the differences that make languages truly unique, as key components of the heritage and as a vital reflection of the communities that use them. This is also a core element of multilingualism in Europe, where all languages are valued as inherent components of the social fabric that connects European citizens in their diversity. The situational context in which the languages are used, which includes societal, economic, educational, and industrial aspects, is incorporated in the DLE definion and metric through the Contextual Factors (CFs), which complement the TFs and are the subject of a companion paper (Grützner-Zahn and Rehm, 2022).

The notion of DLE promoted by ELE does not involve any judgement of the political, social and cultural status or value of the languages, insofar as they collectively contribute to a multilingual Europe, that should be supported and promoted. Alongside the fundamental concept of equality, we also recognise the importance of the notion of equity, meaning that for some languages, and for some needs, a specific effort is necessary. For example, the availability of, and access to, certain services and resources (e. g., to revitalise a language, or to promote the development of education through that language) is very important for some of Europe's languages. With this in mind, the challenge tackled by ELE is to enable all languages of Europe, regardless of their specific circumstances, to realise their full potential, supporting them in achieving full digital equality. The DLE metric, whose TFs are presented here, captures the needs and expectations of the various European languages and the shortfalls with respect to being adequately supported in terms of resources, tools and technological services in the digital age so as to achieve digital language equality.

2.2 Defining the DLE Metric

Following consultations within the ELE consortium, early in the project a definition of DLE was adopted to guide our efforts. The definition of DLE drew inspira-

tion, among others, from the META-NET White Paper Series (Rehm and Uszkoreit, 2012) and from BLARK[3] (Krauwer, 2003), both of which have been used in the past to assess the level of technological support of specific languages. ELE defines DLE as "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; Gaspari et al., 2022). This definition provides the basis to establish a metric that enables the quantification of the level of technological support for each language in scope of ELE with descriptive, diagnostic and predictive value to successfully promote digital language equality. This approach enables comparisons across languages, tracking their advancement towards the goal of DLE, as well as the prioritisation of needs, especially to fill existing gaps, focusing on realistic and feasible targets. The DLE Metric is therefore defined as "a measure that reflects the digital readiness of a language and its contribution to the state of technology-enabled multilingualism, tracking its progress towards the goal of DLE" (Gaspari et al., 2021). The DLE Metric is computed for each language on the basis of various factors, grouped into TFs (e. g., the available language resources, tools and services, which are the focus of this paper) and situational CFs, e. g., societal, economic, educational, industrial, which are described in detail by Grützner-Zahn and Rehm (2022).

2.3 Key Features

The DLE definition and the formulation of the DLE Metric are modular and flexible, i. e., they consist of well-defined separate and independent, but tightly integrated quantifiers, measures and indicators, selected to ensure compatibility and interoperability with the metadata schema adopted by ELE's sibling EU-funded project European Language Grid (ELG)[4] (Labropoulou et al., 2020; Rehm et al., 2020; Rehm, 2022), which plays a crucial technical role with regard to the TFs. ELG maintains a cloud platform that bundles together data sets, corpora, functional software, repositories and applications to benefit European society, industry and academia and administration, while also addressing the fragmentation of the European LT landscape by providing a convenient single access point.

The definitions of DLE and its metric have also been designed to be transparent and similarly intuitive for linguists, LT experts and developers, language activists, advocates of language and human rights, industrial players, policy-makers and European citizens at large, to encourage the widest possible uptake and buy-in. While we wanted them to be founded on solid, widely agreed principles, we also aimed at striking a balance between a methodologically sound and theoretically convincing approach, and a formulation that can be used, among others, to inform future language and LT policies at the local, regional, national and European levels, to guide and prioritise future efforts in the creation, development and improvement of LRs and LTs, with the ultimate goal of achieving DLE in Europe. Through data analytics and visualisation, languages facing similar challenges in this collective endeavour can be grouped together, and requirements can be formulated to support them in remedying the existing gaps and advancing towards full DLE. An analysis of the Technological DLE scores of European languages is presented in Section 3.4.

A crucial feature of the DLE Metric is its dynamic nature, i. e., the fact that its scores can be updated and monitored over time, at regular intervals or whenever one wishes to check the progress or the status of one or more European languages with respect to the goal of achieving DLE. With regard to the TFs, as the ELG Catalogue organically grows over time, the resulting DLE Metric scores will be updated for all European languages, thereby providing an up-to-date and consistent (i. e., comparable) measurement of the level of LT support and provision that each of them has available, also showing where the status is less than ideal or not at the expected level. The DLE Metric can be found, computed dynamically using the data available in the ELG Catalogue, in the ELE/ELG dashboard.[5]

3 Technological Factors

In order to quantify the level of technological support for a language, we consider a set of TFs. Here we briefly describe their main categories, illustrating the breadth and diversity of the LRs and tools that they capture. The first category of TFs includes tools and services that are offered via the web or running in the cloud, but also downloadable tools, source code, etc.; this category encompasses, for example, NLP tools (morphological analysers, part-of-speech taggers, lemmatisers, parsers, etc.); authoring tools (e. g. spelling, grammar and style checkers); services for information retrieval, extraction, and mining, text and speech analytics, machine translation, natural language understanding and generation, speech technologies, conversational systems, etc.

The second category of TFs includes datasets, i. e. corpora or collections of text documents, text segments, audio transcripts, audio and video recordings, etc., monolingual or bi-/multilingual, raw or annotated. It also encompasses language models and computational grammars and lexical and conceptual resources, including resources organised on the basis of lexical or conceptual entries (lexical items, terms, concepts, etc.) with their supplementary information (e. g., grammatical, semantic, statistical information, etc.), such as computational lexica, gazetteers, ontologies, term lists, thesauri, etc.

[3] http://www.blark.org

[4] https://live.european-language-grid.eu

[5] https://live.european-language-grid.eu/catalogue/dashboard

The technological component of the DLE Metric and the resulting Technological DLE score per language are based on the number of LRs available for a given language. Although an essential aspect of a language's digital readiness is the number of available LRs, equally important are the types and features of these LRs, insofar as they indicate how well a language is supported in all different LT areas. To capture such aspects with the DLE metric, in addition to raw counts of available LRs, the following features of LRs have also been taken into account:

- resource type
- resource subclass
- linguality type
- media type covered or supported
- annotation type, where relevant
- domain covered, where relevant
- function/task performed (for tools/services only)
- conditions of use

The values of these features are appropriately weighted to contribute to the resulting Technological DLE score. The weights applied to LR feature values are listed in Tables 1 and 2 in the Appendix and further discussed in Section 3.1.

3.1 Applying Weights to the Factors

The weights are applied to LR feature values, in order to reward the contribution of a LR to DLE with regard to the relevant TFs. This is based on the assumption that some LR features contribute more effectively to achieving DLE than others. Higher weights are assigned to feature values related to (i) more complex technologies, e. g., LTs that employ or support more than one modality, (ii) more "expensive" datasets/tools, in terms of the investment required to build them, (iii) more "open" or freely available datasets and tools, and (iv) additional or broader envisaged applications.

One guiding consideration in developing the DLE Metric, and especially in assigning the weights of the features and their values for the TFs, was to make the fewest possible assumptions about the (preferred) end-uses and actual application scenarios that may be most relevant to users. These inevitably vary widely due to a number of variables that are impossible to establish *a priori*. We therefore refrained from predetermining particular preferred end-uses when proposing the full specification of the DLE Metric, which otherwise would risk it being unsuitable for some end-users and applications. In Tables 1 and 2 in the Appendix we present the TFs of the DLE Metric with their weights; this set-up is subject to revision as more experiments are run within ELE in addition to those reported in Section 3.2 to adjust the weights, so that the Technological DLE scores capture and reflect fairly the actual level of LT support for the ELE languages.

The features and values for the LRs and LTs that make up the TFs are derived from the metadata schema used in the ELG Catalogue (Labropoulou et al., 2020; Rehm et al., 2020); the weights assigned to them are listed in Table 1 and Table 2 in the Appendix for LRs and tools, respectively. Here we briefly review some of the key features of the TFs, focusing on those that can have several values, which are of particular interest because they show the level of detail and granularity of the metadata accompanying the records included in the ELG Catalogue.

A varied feature within LRs is that of "Annotation Type", which has many possible values. For the first implementation of the DLE Metric, we have assigned a constant very small fixed weight, also based on the fact that some LRs can possess several annotation types. A similar consideration applies to the "Domain" feature, which has many possible values for LRs and for tools: in these cases, the weights assigned to "Domain" values in the first instance are fixed and relatively small, again considering that multiple domains can be combined in a single LR or tool. In addition to "Domain", another feature that appears both in LRs and tools is "Conditions of use"; the weights proposed for this feature of the TFs are identical for the corresponding values of "Conditions of use" across datasets and tools. In the case of (much) more restrictive licensing terms, lower weights are assigned than to liberal use conditions, so they contribute (much) less to the Technological DLE score for the LR in question, and therefore to the cumulative DLE Metric score for that language.

3.2 Experiments with ELG

To experiment with different set-ups for the TFs of the DLE Metric, we used the Catalogue of the European Language Grid, which in early 2022 contained approx. 11,500 records, out of which about 75% were datasets and resources (corpora, lexical resources, models and grammars) and the rest were tools and services, covering almost all European languages. These records contain multiple levels of metadata granularity. We consider the current status of the ELG repository to be representative with regard to the current existence of LT resources for Europe's languages, so it is used by ELE as its empirical basis for the computation of the technological DLE score.

The ELG Catalogue includes metadata of both LRs and LTs for all ELE languages. Each resource and tool has several features and associated values, as shown in the Appendix. Each feature was assigned a weight to calculate the Technological DLE score on a per-language basis, comparing the resulting scores of a number of alternative set-ups, considering especially where each language stood in relation to all the others and how their relative positioning changed as a result of assigning different weights to the various feature values. This was an efficient and effective method to gradually refine the set-up of the TFs and propose the implementation of the relevant weights.

The experiments have shown that the global picture of

the DLE Metric scores for the languages targeted by ELE tends not to change dramatically as the weights assigned to the feature values vary. We have experimented both with very moderate and narrow ranges of weights, and with more extreme and differentiated weighting schemes. Since, ultimately, any changes are applied across the board to all LRs and tools included in ELG for all languages, any resulting changes propagate proportionally to the entire set of languages, thus making any dramatic changes rather unlikely, unless one studiously unduly rewards (i. e., games) specific features that are known to disproportionately affect one or more particular languages. It should immediately be clear that this would be a biased and unfair application of the DLE Metric, and should be avoided at all costs.

Our experiments demonstrate that the overall representation of the languages tends to be relatively stable. This is due partly to the sheer amount of features and possible feature values that make up the TFs. As a result, even if one changes the weights, with the exception of minor and local fluctuations, three main phenomena are generally observed: (i) the overall relative positioning of the languages remains largely stable, with a handful of languages standing out with the highest Technological DLE scores (English leading typically over German, Spanish and French, with the second language having roughly half the Technological DLE score of English), the minimally supported languages still displaying very low scores, and a substantial group of evenly distributed languages towards the middle; (ii) clusters of languages with similar LT support according to intuition and expert opinion remain ranked closely together, regardless of the adjustments made to specific weights for individual features and their values; and, finally, (iii) even when two similarly supported languages change relative positions (i. e., language A overtakes language B in terms of Technological DLE score) as a result of adjusting the weights assigned to features and their values, their absolute Technological DLE scores remain very close.

We have also performed focused checks on pairs or small sets of languages spoken by comparable communities and used in similar circumstances, and whose relative status in terms of LT support is well known to the experts. These focused checks have involved, e. g., Basque and Galician, Irish with respect to Welsh, and the dozen local languages of Italy (also with respect to Italian itself), etc. Overall, the general stability and consistency demonstrated by the Technological DLE scores across different set-ups of weight assignments for the various features and their possible values for TFs provides evidence of its validity as an effective tool to guide developments and track progress towards full DLE for all of Europe's languages by 2030.

Table 1 and Table 2 in the Appendix provide the configuration of the weights assigned to the TFs to compute the Technological DLE score. This set-up is subject to adjustments as more experiments are conducted to check any need to refine the weights, in the interest of making the DLE Metric truly representative of the actual level of LT support for European languages. This approach will ensure that the DLE metric optimally captures the real situation and also effectively reflects the needs and aspirations of all of Europe's languages and their communities for the future in the digital age.

3.3 DLE Metric Formula

Based on the above, the steps to calculate the Technological DLE score are as follows:

1. Each LR in the ELG Catalogue (dataset or tool) obtains a score ($Score_{LR}$), which is equal to the sum of the weights of its relevant features. Specifically for features Annotation Type and Domain, instead of simply adding the respective weight, the weight is multiplied by the number of unique feature values possessed by the LR in question.

   Example: Suppose an LR in the ELG catalogue (LR1) has the following features: corpus, annotated, monolingual, with three different annotation types (morphology, syntax, semantics), with text as media type, covering one domain (e. g. finance), with conditions of use research use allowed. Then, using the weights proposed in Tables 1 and 2 in the Appendix, LR1 is assigned the following score:

   $Score_{LR1} = 5 + 1 + 2.5 + (3 * 0.25) + 1 + (1 * 0.3) + 3.5 = 14.05$

2. To compute the Technological DLE score for language X ($TechDLE_{LangX}$), for all LRs that support language X (LR1, LR2, ... LRN), one sums up the $Score_{LR}$ of all LRs that support language X (LR1, LR2, ... LRN), i. e.

   $TechDLE_{LangX} = Score_{LR1} + Score_{LR2} + \ldots + Score_{LRN}$

3.4 ELE Languages: Technological DLE Scores

Based on the weights, the Technological DLE scores of Europe's languages as of mid-May 2022 are presented in Figure 1. To allow for a more fine-grained visual representation, Figures 2 and 3 in the Appendix show the first and the second half of the languages, respectively, using more appropriate scaling of the score ranges.

Not surprisingly, English is by far the most well-resourced language of Europe, leading the way over German and Spanish, that follow with very similar Technological DLE scores, which are roughly half that of English. French is at present the fourth most well-resourced European language. Finnish, Italian and Portuguese follow at some distance, and it is interesting to note that the next cluster of languages that are spoken by sizeable communities in Europe (Polish, Dutch and Swedish), still in the top ten of the overall list of languages, have a Technological DLE score that is roughly six times lower than that of English.

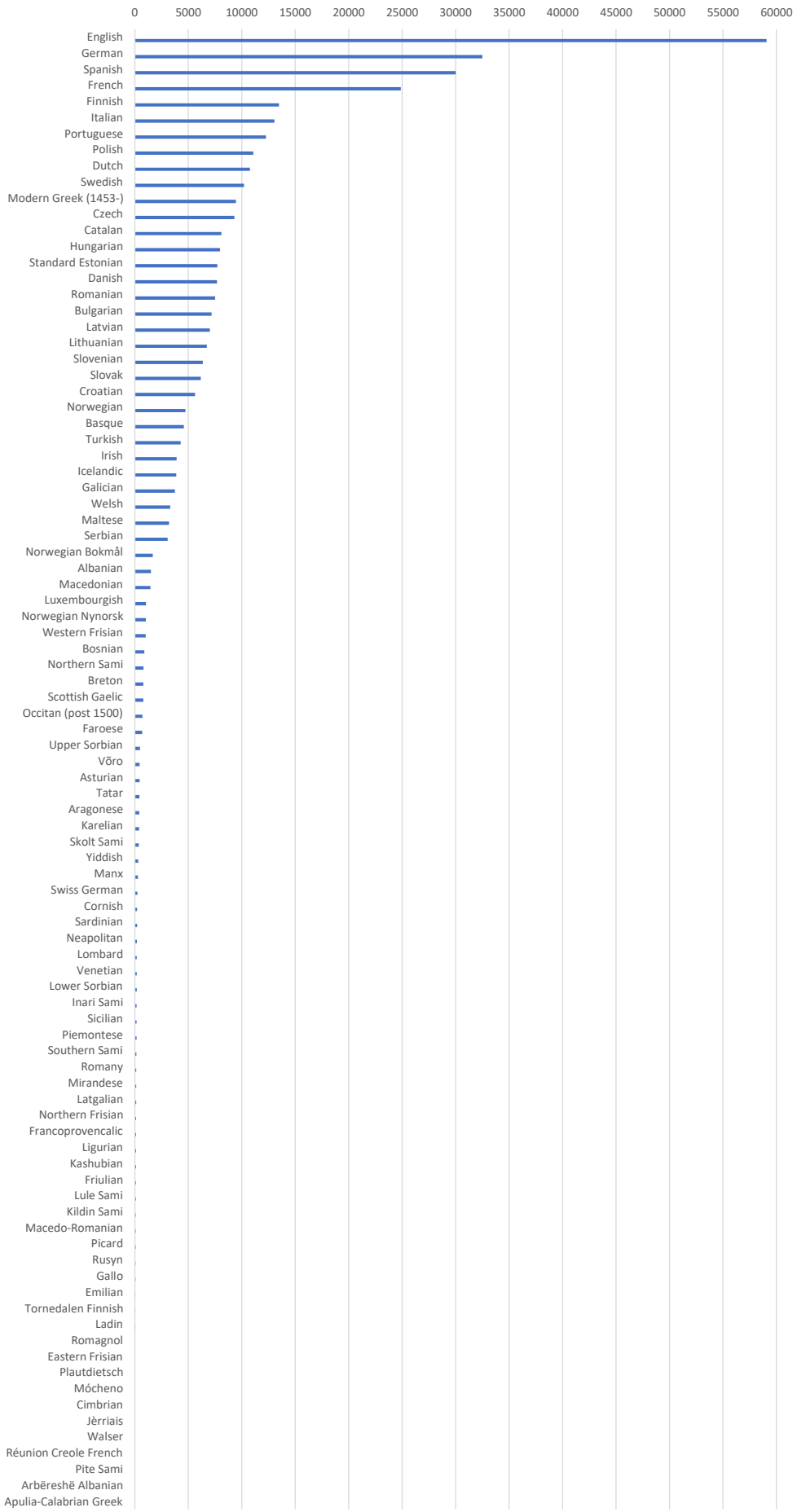

Figure 1: Technological DLE scores for all ELE languages as of mid-May 2022.

A number of observations can be made on the basis of the Technological DLE scores shown in Figures 1, 2 and 3: first, one can see that even some official EU and national languages are not particularly well-supported, at least in comparison with the leading languages, first and foremost English. It should also be noted that some non-official EU languages such as Catalan, Basque, Galician and Welsh appear to be relatively well-supported, also in comparison with some official EU languages. In addition, it is quite striking that several European languages currently represented in ELG have very low Technological DLE scores, which points to the fact that currently most of them have hardly any datasets and basic LTs that are essential for them to remain alive and be used by the respective communities, so as to prosper in the digital area.

3.5 Open Issues and Challenges

The Technological DLE scores discussed here do not take into account the size of LRs or the quality of LRs and LTs. While these are important features, there exist a large variety of size units for LRs, and the way for measuring data size is not standardised, especially for new types of LRs such as models. Regarding the quality of tools in particular, while some information on the Technology Readiness Level scale is available in the ELG Catalogue, the large number of null values does not make it possible to take this aspect into account at the moment. These are shortcomings that we intend to revisit in subsequent efforts, in order to overcome these limitations and improve the overall accuracy and granularity of the Technological DLE score.

As far as datasets are concerned, there could be benefits to setting a minimum size criterion to include LRs such as corpora or grammars in the computation of the Technological DLE score, e. g., to avoid using small resources that cannot be realistically applied in technology development scenarios. However, at present it would be difficult to establish arbitrarily what this minimum size threshold should be, also in recognition of the specifics of the several languages covered by ELE. As a result, the decision was made not to set any minimum size requirement for LRs. The thinking behind this choice was that relatively small data sets are common in less-resourced languages, for particular domains, etc., and there is the possibility to merge small data sets to create bigger ones that would, in fact, be useful, e. g., in domain adaptation for machine translation. More broadly, ELE intends to foster a culture of valuing all and any LRs, especially for less-resourced languages, judiciously balancing the importance given to the size, quantity, diversity and quality of the LRs.

Finally, projects and organisations are not taken into account for the time being, partly due to the difficulty of attributing them specifically to individual languages, even though the possibility remains open to include these additional features and values in the computation of the Technological DLE score at a later stage.

4 Conclusions and Future Work

We introduce the notion of DLE and describe the DLE Metric, focusing in particular on the Technological DLE score, as developed in the ELE project. By providing an empirically-grounded and realistic quantification of the level of technological support possessed by the languages of Europe, the DLE Metric, whose TFs are complemented by the CFs, will contribute to the formulation of the sustainable evidence-based SRIA and Roadmap that will drive future efforts in equipping all European languages with the LTs needed to achieve full DLE in Europe by 2030; the DLE Metric will also provide a transparent means to track and monitor the actual progress in this direction.

With regard to the TFs, the close collaboration with the sister project ELG has been particularly valuable, in that the TFs rely on the metadata in the ELG Catalogue as the ground truth and empirical foundation to measure and quantify the level of digital readiness of the languages covered by ELE. The overview of the TFs is accompanied by an in-depth discussion of the scoring and weighting mechanism adopted for the computation of the Technological DLE score, that is illustrated to explain the overall design of the features and values that contribute to the TFs.

The weights assigned to the features to compute the Technological DLE score can be adjusted going forward. This approach would be useful to address developments ensuing from advances made in LT and as new paradigms or technologies become the state of the art, potentially also as new types of resources emerge and are recognised as crucial for LT support. The ELE consortium views the DLE Metric as a flexible tool, with the possibility of updating and revising if and as needed the exact configuration of the TFs and CFs.

We are confident that the concept of DLE and its associated Metric described here represent valuable tools on which to base subsequent efforts to measure and improve the readiness of Europe's languages for the digital age, also taking into account the situational contexts in which the languages are used via the CFs. By drawing on the descriptive, diagnostic and predictive value of the DLE Metric, the community will have a solid and verifiable means of pursuing and evaluating much-needed developments in the interest of all European citizens. In conclusion, we hope that the DLE Metric will be recognised as a helpful tool by a range of stakeholders at various local, regional, national and European levels who are committed to preventing the extinction of European languages under threat, and who are interested in promoting their prosperity. Such stakeholders include decision- and policy-makers, industry leaders, researchers, developers, and citizens across Europe who will drive forward future developments in the fields of LT and language-centric AI.

Acknowledgements

The work presented in this article was co-financed by the European Union under grant agreement no. LC-01641480 – 101018166. Part of the work has also been supported by the ADAPT Centre for Digital Content Technology which is funded under the Science Foundation Ireland (SFI) Research Centres Programme (Grant No. 13/RC/2106) and is co-funded under the European Regional Development Fund.

5 Bibliographical References

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

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

European Parliament. (2018). Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI). http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf.

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

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. Last accessed: 08.02.2022.

Gaspari, F., Grützner-Zahn, A., Rehm, G., Gallagher, O., Giagkou, M., Piperidis, S., and Way, A. (2022). D1.3 Digital Language Equality (full specification of the concept).

Grützner-Zahn, A. and Rehm, G. (2022). Introducing the Digital Language Equality Metric: Contextual Factors. In Itziar Aldabe, et al., editors, *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*, Marseille, France.

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, July. 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.

Labropoulou, P., Gkirtzou, K., Gavriilidou, M., Deligiannis, M., Galanis, D., Piperidis, S., Rehm, G., Berger, M., Mapelli, V., Rigault, M., Arranz, V., Choukri, K., Backfried, G., Pérez, J. M. G., and Garcia-Silva, A. (2020). Making Metadata Fit for Next Generation Language Technology Platforms: The Metadata Schema of the European Language Grid. In Nicoletta Calzolari, et al., editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3421–3430, Marseille, France. European Language Resources Association (ELRA).

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

Georg Rehm et al., editors. (2022). *European Language Equality: A Strategic Agenda for Digital Language Equality*. Cognitive Technologies. Springer. Forthcoming.

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

Rehm, G., Gaspari, F., Rigau, G., Giagkou, M., Piperidis, S., Resende, N., Hajic, J., and Way, A. (2022). The European Language Equality Project: Enabling Digital Language Equality for all European Languages by 2030. *EFNIL Conference Publications Cavtat 2021*. 23 pp.

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

Appendix

Feature	Feature Value	Weight
Resource Type	corpus	5
	lexical conceptual resource	1.5
	language description	3.5
Subclass	raw corpus	0.1
	annotated corpus	2.5
	computational lexicon	2
	morphological lexicon	3
	terminological resource	3.5
	Wordnet	4
	Framenet	4
	model	5
	each of the others (there are 15 more)	0.5
Linguality Type	multilingual	5
	bilingual	2
	monolingual	1
Media Type	text	1
	image	3
	video	5
	audio	2.5
	numerical text	1.75
Annotation Type	*each of these – can be combined in a single LR*	0.25
Domain	*each of these – can be combined in a single LR*	0.3
Conditions of Use	other specific restrictions	0.5
	commercial uses not allowed	1
	no conditions	5
	derivatives not allowed	1.5
	redistribution not allowed	2
	research use allowed	3.5

Table 1: Weights assigned to the Technological Factors of the DLE Metric – Language Resources.

Feature	Feature Value	Weight
Language Independent	false	5
	true	1
Input Type	input text	2
	input audio	5
	input image	7.5
	input video	10
	input numerical text	2.5
Output Type	output text	2
	output audio	5
	output video	10
	output image	7.5
	output numerical text	2.5
Function Type	text processing	3
	speech processing	10
	information extraction and information retrieval	7.5
	translation technologies	12
	human-computer interaction	15
	natural language generation	20
	support operation	1
	image/video processing	13
	other	1
	unspecified	1
Domain	*each of these – can be combined in a single tool*	0.5
Conditions of Use	unspecified	0
	other specific restrictions	0.5
	no conditions	5
	commercial uses not allowed	1
	derivatives not allowed	1.5
	redistribution not allowed	2
	research use allowed	3.5

Table 2: Weights assigned to the Technological Factors of the DLE Metric – Tools.

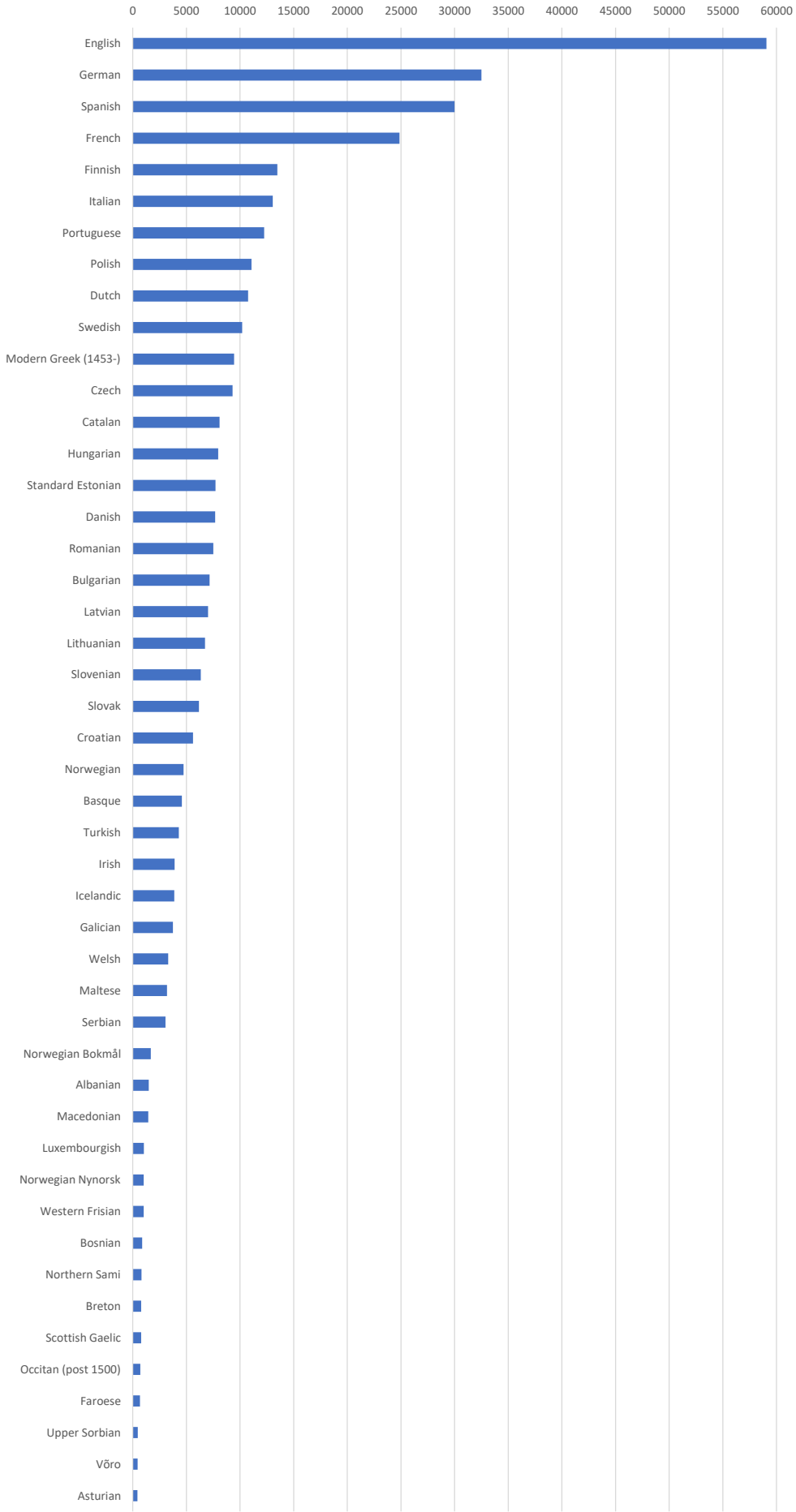

Figure 2: First half of the languages listed in Figure 1 on a range of 0-60,000 Technological DLE score points.

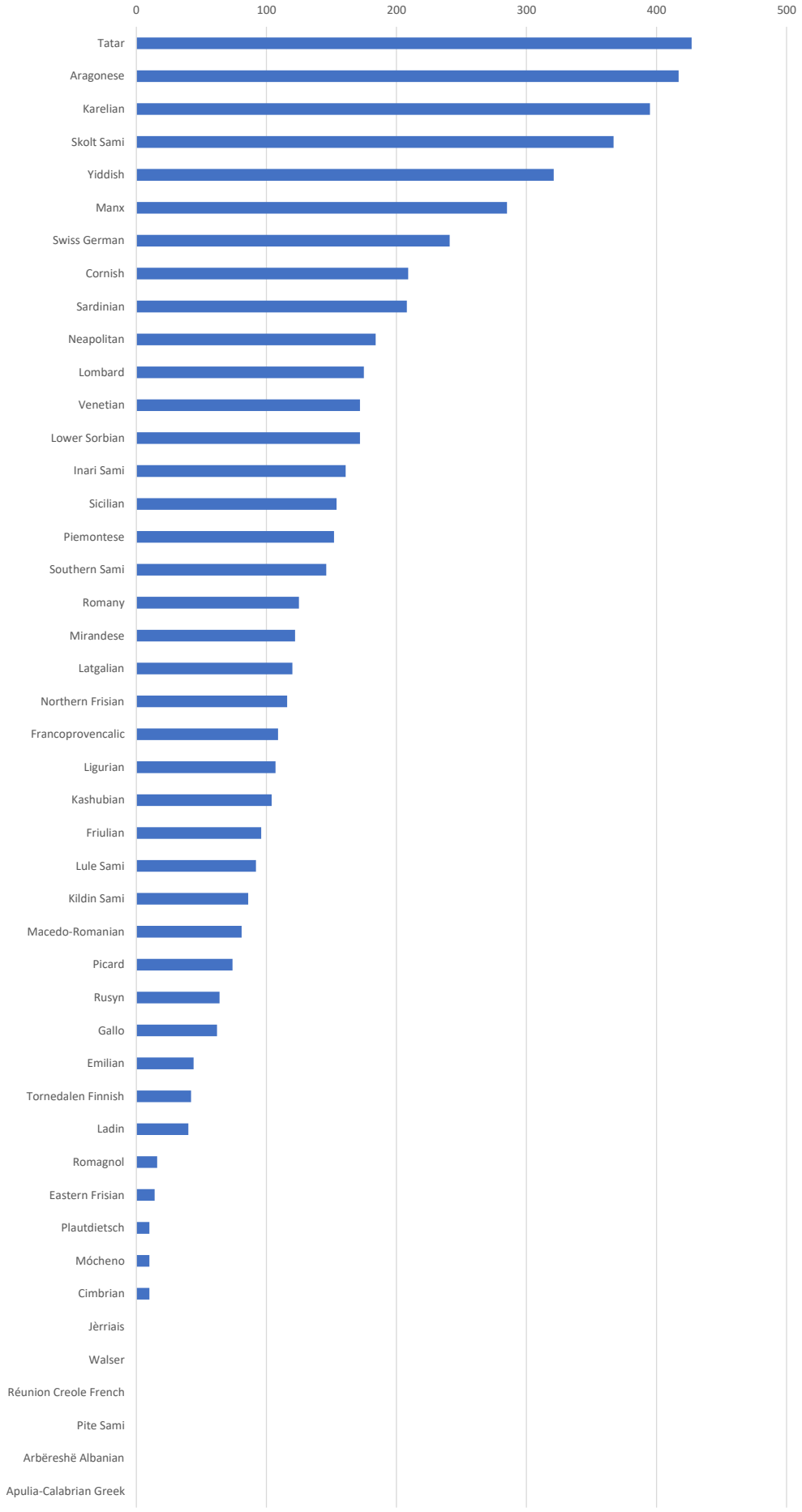

Figure 3: Second half of the languages listed in Figure 2 on a range of 0-500 Technological DLE score points.