Proceedings of the TDLE Workshop @LREC2022, pages 36–45
Marseille, 20 June 2022

Measuring HLT Research Equality of European Languages

Gorka Artola, German Rigau
HiTZ Basque Center for Language Technologies - Ixa
University of the Basque Country UPV/EHU
{gorka.artola, german.rigau}@ehu.eus

Abstract

This work explores quantitative indicators that could potentially measure the equality and inequality research levels among the languages of the European Union in the field of human language technologies (HLT research equality). Our ultimate goal is to investigate European language equality in HLT research considering the number of papers published on several HLT research venues that mention each language with respect to their estimated number of speakers. This way, inequalities affecting HLT research in Europe will depend on other factors such as history, political status, GDP, level of social or technological development, etc. We have identified several groups of EU languages in the proposed measurement of HLT research equality, each group comprising languages with large differences in the number of speakers. We have discovered a relative equality among surprisingly different languages in terms of number of speakers and also reAll data and code will be released upon acceptance.

Keywords: human language technologies, equality, European languages

1. Introduction

The language landscape in the European Union (EU) comprises 24 official EU Member State languages, including three different alphabets, and more than 60 regional and minority languages (Pastor, 2018), including languages of relevant trade partners and immigrant communities. The fact that several of the regional languages enjoy the same level of official status as the corresponding EU Member State language in their respective regions, e.g., Aranese, Basque, Catalan, Galician, Luxembourgish, Scottish Gaelic and Welsh, and also the fact that different levels of protection by local authorities have been developed across Europe for several non-official regional or minority languages, are both European particularities not easily found in other societies in the world. One of the reasons for this diversity and public support is that multilingualism is one of the core values of the EU based on the motto 'United in diversity', and a matter deeply embedded even in the most basic regulation of the EU. A remarkable example of this can be seen in the Article 165(2) of the Treaty on the Functioning of the EU (TFEU)[1], which emphasises that *Union action shall be aimed at developing the European dimension in education, particularly through the teaching and dissemination of the languages of the Member States, while fully respecting cultural and linguistic diversity (Article 165(1) TFEU).* Thus, for instance, the EU works with Member States to protect minorities, on the basis of the Council of Europe's European Charter for Regional or Minority Languages[2], or to promote multilingualism in the development of the EU Digital Single Market. The EU resolution "Regional and lesser-used languages - enlargement and cultural diversity"[3] is another relevant example of the subject.

A wide diversity of languages in Europe are expected to coexist, interact and evolve efficiently as equals. The strength of the multilingual EU is therefore believed to be based on the equality among European languages, but protecting and promoting language diversity, and gaining as a consequence a recognisable equality among languages operating simultaneously in a society is not an easy endeavour. The challenge is even more complex when, like in the case of the EU, the society is a conglomerate of smaller regional societal bodies with high levels of interaction and interdependence among them, but each one with a different profile and mix of coexisting languages.

The sources of inequalities among languages are multiple and possibly related to almost any dimension of its human and social condition. Economy, demography, history, geography, religion, policy and a long *etcetera* shape each and every language, making their comparison very complex. Language equality is a vibrant and remarkable challenge, and a research field that is building its own foundations. This work intends to contribute to both the European challenge and the emerging research field through the deliberation about the equality of European languages in their digital facet, particularly in the field of research in Human Language Technologies (HLT research equality).

In addition, the HLT community is currently developing powerful new deep learning techniques and tools that are revolutionizing the approach to HLT tasks. We are gradually moving from a methodology in which a pipeline of multiple modules was the typical way to im-

[1] `http://data.europa.eu/eli/treaty/tfeu_2012/oj`

[2] `https://www.coe.int/en/web/european-charter-regional-or-minority-languages`

[3] `https://www.europarl.europa.eu/doceo/document/TA-5-2003-0372_EN.html`

plement HLT solutions, to architectures based on complex neural networks trained with vast amounts of text data. The success in HLT has been possible because of the confluence of four different research trends: 1) mature deep neural network technology, 2) large amounts of data (and for NLP processing large and diverse multilingual textual data), 3) increase in High Performance Computing (HPC) power in the form of GPUs, and 4) application of simple but effective self-learning approaches. Interestingly, the application of zero-shot to few-shot transfer learning with multilingual pretrained language models, prompt learning and self-supervised systems opens up the way to leverage HLT for less developed languages (Goodfellow et al., 2016; Devlin et al., 2019; Liu et al., 2020; Torfi et al., 2020; Wolf et al., 2020). However, a growing concern is that due to unequal access to these resources only certain IT companies and elite universities have advantages in modern HLT research (Ahmed and Wahed, 2020).

After this introduction, Section 2 presents several studies carried out on language equality. Sections 3 and 4 describe our research framework and Section 5 provides an in-depth analysis of the HLT research equality of the European languages on the basis of the quantitative indicators proposed in this work. Finally, Section 6 summarizes our main findings and presents our future work.

2. Related work

Given the role of HLT in everyone's daily lives, many expert practitioners are directly concerned by language diversity in HLT research and development.[4] For instance, Sayers et al. (2021) emphasise a range of groups who will be disadvantaged. Looking ahead, they see many intriguing opportunities and new capabilities, but also a range of other sources of uncertainties and inequalities. Joshi et al. (2020) examine the relation between the types of languages, resources and their representation in NLP conferences over time. As expected, only a very small number of the over 7000 languages of the world are represented in the rapidly evolving HLT field. Just a handful of languages are covered by current NLP systems, drawn from a few dominant language families. As a result, most linguistic phenomena from typologically diverse languages have never been incorporated to our HLT research (Ponti et al., 2019). Blasi et al. (2021) study the systematic inequalities in HLT across World languages. After English, a handful of Western European Languages dominate the field -in particular German, Spanish and French- as well as even fewer non-Indo-European languages, primarily Chinese, Japanese and Arabic. This investigation suggests that it is the economy of the users of a language (rather than demography) what drives the development of HLT. While language diversity is at the core of Europe identity and multilingual society, many of our languages are in danger of digital extinction because they are not sufficiently supported through HLT (Moseley, 2010). The EUROMAP Language Technologies was the first project investigating the state-of-the-art of HLT research and take-up in Europe, as well as the background situation in each country (Joscelyne and Lockwood, 2003). *META-NET White Paper Series: Europe's Languages in the Digital Age* (Rehm and Uszkoreit, 2012; Rehm et al., 2014) provide the first systematic study about the technology support of Europe's languages. The Rehm and Hegele (2018) survey represents the voices of more than 600 respondents from more than 50 countries working on LT. Rehm et al. (2020) present an overview of various European HLT and AI reports, and perform an extensive qualitative analysis of the landscape of research on HLT research in all the Member countries of the EU.

Both works of Joshi et al. (2020) and Blasi et al. (2021) consider and use in their studies the number of papers mentioning each language as an element, among many others, to measure inequalities in HLT. Both conclude that the main European languages are among the most equal and best represented languages in HLT, considering the large-grained scope of the 7,000 estimated languages in the world. Our work intends to explore the potential of simple indicators based also on the numbers of papers mentioning each language to measure fine-grained inequalities in HLT research, and complement with quantitative data the qualitative study on European HLT research of Rehm et al. (2020). We believe that this approach could unveil inequalities not easily demonstrable by other means, that are undermining the European language diversity protection goal, and will help identify relatively low-resourced and endangered languages in HLT research even within the theoretically strongest ones.

The work in progress in the European Language Equality Project (ELE)[5] is also worth to be noted. With a large and all-encompassing consortium consisting of 52 partners covering all EU Countries, research and industry and all major pan-European initiatives, ELE develops a strategic research, innovation and implementation agenda as well as a roadmap for achieving full digital language equality in Europe by 2030.

3. Initial hypothesis

Research, development and innovation in HLT is, generally, affordable and accessible for societies that have reached certain level of human and economic development. This is believed to be the case of the Countries and Regions comprising the EU, and together with the recognition and protection levels that the EU and member states offer to the variety of European languages

[4] https://gitlab.com/ceramisch/eacl21diversity/-/wikis/EACL-2021-language-diversity-panel

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

creates a unique case of theoretical favourable environment for equality among these different languages.

The initial hypothesis of this work is that, particularly in the field of HLT research, the languages of the EU should show a relevant degree of equality and that any inequality must respond to other factors than technological, social, cultural or regulatory barriers. The identification of the eventual inequality among European languages in this field may lead to effective direct intervention by the stakeholders (policy makers, academy, industry and any other) that could have legitimate interest in correcting the divergence. Also, on the other hand, it could confirm the effectiveness of existing scientific, regulatory, policy and societal dynamics in the purpose of achieving the language equality.

Finally, the focus of our study in HLT research is expected to be further beneficial contributing to the general goal of language equality, provided these technologies have precisely the ability to potentially reduce inequalities among languages through the use of digital technologies. An endangered language, or a language not reaching sufficient equality with others, may converge faster to equality taking advantage of HLT research, but failing or performing poorly on it may be an unbridgeable barrier to gain overall language equality, or even a menace towards the ongoing digital transformation.

4. Selected Languages, Data Sources and Measurement Indicators

For the identification, denomination and basic characterisation of European Languages involved in the study, and also for the estimation of the number of speakers in Europe for each language, we have followed the criteria designed by the previously mentioned European Language Equality Project (ELE). The selection of the source of data itself introduces a certain degree of a bias, particularly in non-official languages or in cases of very few speakers, on which there is no consensus denominating the language or the speaker statistics, and this will be taken into account in the analysis of the results.

We make the working assumption that a mention of a language in a research paper likely entails that the underlying research involves in some extent this language, that the more the papers mentioning a particular language the more the chance that HLT research is having a positive impact on that language, and the better is its position in the field of HLT research. Of course, we do not pretend this to be a measure of the overall HLT equality between languages, but just a measure of the presence of each language in HLT research.

The first basic indicator we have selected to explore the quantitative measurement the equality among languages in the field of HLT research is the number of scientific documents that mention each language published in the period from 2000 to 2020. We will refer to this measurement as the *absolute metric*. Not being feasible to gather and analyse the whole global scientific production in this field, we have selected a group of relevant venues and sources where the most relevant scientific documents of the field are most likely to have been published. These selected sources are the Proceedings of the bi-anual Language Resources and Evaluation Conference (LREC)[6], the Annual Meeting of the Association for Computational Linguistics (ACL)[7], the Conference on Empirical Methods in Natural Language Processing (EMNLP)[8], and the Computational Linguistics Journal (CL)[9]. The selection of these publication venues also introduces a bias to be taken into account in any analysis of these measurements. We have crawled all documents in pdf format published in these venues from 2000 to 2020 available in the ACL Anthology website[10], computationally extracted the text of these files transforming them in plain text files, and found what EU languages are mentioned in each document, according to the list developed by the ELE project[11]. Proper names that are the same as EU languages but not refer to a Language, e.g. ”Basque” in the name ”University of the Basque Country”, have also been detected and not included in the counts of language mentions. Table 1 shows the number of research papers processed from each source.

Source	Papers
LREC	7,175
ACL	9,672
EMNLP	7,087
CL	1,977
Total	25,911

Table 1: Number of processed research papers per source

As a second quantitative measurement of HLT research equality, we propose to compare also the number of documents mentioning each language per million of speakers. We will refer to this indicator as the *relative metric*. The rationale behind the proposal of this indicator is an attempt to remove from the analysis the effect that plain demography may have in HLT research. Between two hypothetical languages where all variables affecting them could be considered exactly the same with the exception of the number of speakers, it would be reasonable to expect to have more researchers in HLT in the most spoken one of them, and also more likely that they mention their own language in the scientific production. Thus, in the extent the *ab-*

[6] https://aclanthology.org/venues/lrec/

[7] https://aclanthology.org/venues/acl/

[8] https://aclanthology.org/venues/emnlp/

[9] https://aclanthology.org/venues/cl/

[10] https://aclanthology.org/

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

solute metric has no capacity to give information about this subject we have considered the need to introduce this second metric.

5. Analysis of language equality

As a starting reference point, Figure 1 describes, the breakdown of the number of estimated speakers in the EU for the languages of the EU considered in the ELE project sorted by the share of each language in the total. Around 80% of speakers are concentrated in 8 languages out of 67 main EU languages. This top group includes three of what we could define as ”global” languages, English, Spanish and Portuguese, languages born in Europe but with more speakers abroad than in their countries of origin. Similarly, 75% of the speakers concerned by the top 8 languages are concentrated in only three languages (English, German, French). Considering only this demographic metric, languages of the EU are inherently and deeply non equal.

Figure 2 shows the breakdown of European languages sorted by total number of documents mentioning each language in the sources selected for the study. If we take this *absolute metric* as a measurement of the HLT research equality of European languages, this figure shows a high degree of overall inequality in this field, but comparing this figure with Figure 1 we may consider groups of languages with some extent of equality on HLT research within the global intrinsic inequality. English grows remarkably, comparatively to all the rest in this metric, but German and French seem to reduce and appear closer to the position of Spanish and Italian. Similarly, Dutch, Czech, Swedish, Portuguese and Turkish also grow with respect to their relative sizes in Figure 1. Maybe the most remarkable advancements in ranking are those of Turkish and Portuguese, languages that like English are, in addition to European Languages, National Official languages of very large countries outside Europe like Turkey, Brazil, etc. We can also observe that, while Greek seems to maintain its position, other strong languages in Europe like Romanian, Hungarian and Polish in terms of number of speaker loose ground compared to less spoken languages. These variations in the relative position of each language in these rankings suggest that there could be HLT research equality and inequality clusters of different nature among European languages, not affecting only low-resource and endangered languages but also some of the most spoken languages and National Official languages in Europe.

Figure 3 shows the breakdown of the number of documents mentioning each language per million of speakers of that language in Europe. We have removed from this ranking languages below 100.000 speakers to avoid introducing non representative distortions in the comparison with languages with several millions of speakers. Observing the pie chart, and comparing it to the ones in figures 1 and 2, we can observe that, according to this *relative metric*, the differences between languages are lower showing higher overall HLT research equality levels among EU languages. At a first glance, now the most spoken and most mentioned languages rank in middle to lower positions in the list, and on the contrary, some languages with lower numbers of speakers like Basque, Icelandic and Breton rise to the top of the list. Remarkably, Turkish also appears in top position despite being a language received in Europe through immigration. But also in this case, we can observe different circumstances among languages. With this metric we can observe HLT research inequalities within the group of less spoken, potentially endangered languages. We can observe some of these languages in

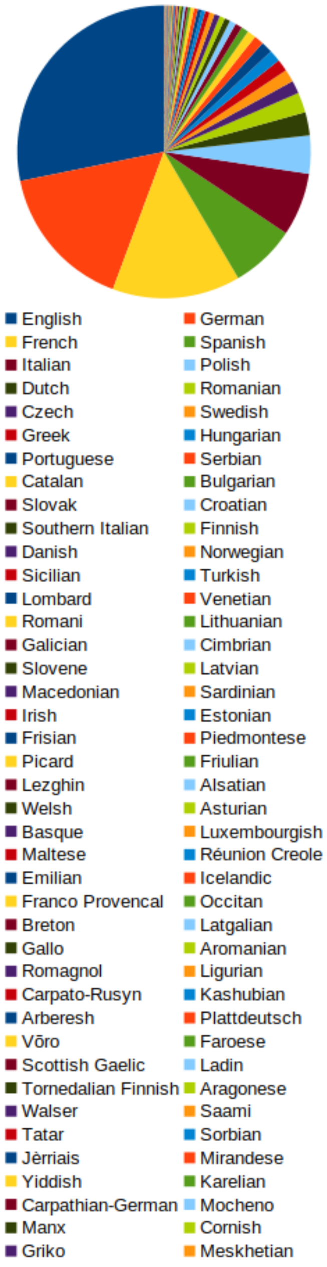

Figure 1: Proportion of speakers in the EU per language of the EU.

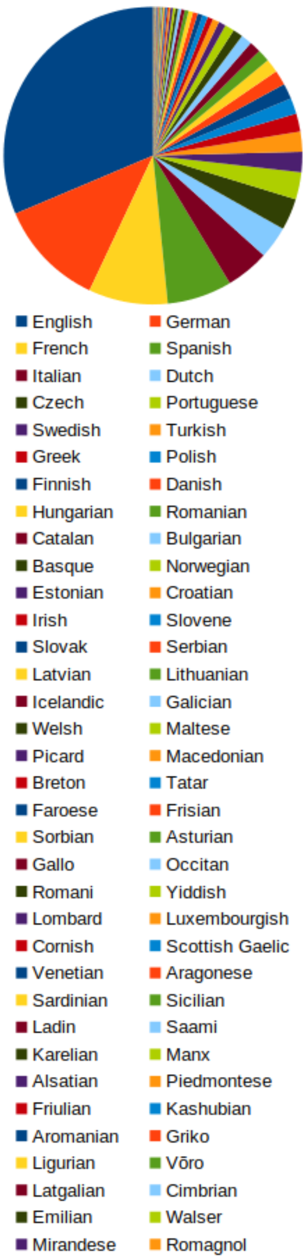

Figure 2: Proportion of documents mentioning languages of the EU (only languages with published documents).

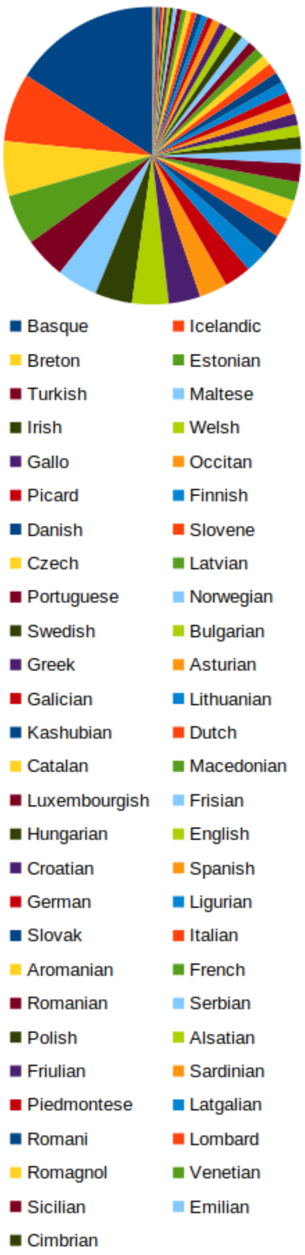

Figure 3: Proportion of documents mentioning languages of the EU per million of speakers (only languages with published documents and with over 100.000 speakers in the EU).

the top positions in the chart, and also some of them in the lowest positions, evidencing a particular kind of inequality in HLT research in European languages, the ones that rank in lower positions both in the *absolute metric* and the *relative metric*.

Table 2 includes the EU languages identified in the ELE project for which no mentions have been found in the HLT research publications. We find in this table eight languages classified by the ELE project as Additional Languages and four Endangered Languages spoken in Europe, and none of them happens to enjoy any officially recognised status by the regional governments of the areas where they are spoken.[12] The presence of Southern Italian, with 5,700,000 estimated speakers, and less spoken but still relevant languages like Lezghin and Réunion Creole in this list suggests existence of weaknesses of some nature around these languages and HLT research. Anyhow, this list brings to surface the potential existence of a group of EU lan-

[12] It is also possible that some research papers identify these languages with other names than the ones given in the ELE project, or that HLT research on these languages is published on venues not included in this study.

ELE language	ELE Classification	Speakers
Southern Italian	Additional Languages spoken in Europe	5,700,000
Lezghin	Additional Languages spoken in Europe	600,000
Réunion Creole	Additional Languages spoken in Europe	484,000
Franco Provencal	Endangered Languages spoken in Europe	227,000
Carpato-Rusyn	Additional Languages spoken in Europe	135,810
Arberesh	Endangered Languages spoken in Europe	100,000
Plattdeutsch	Additional Languages spoken in Europe	90,000
Tornedalian Finnish	Additional Languages spoken in Europe	30,000
Jèrriais	Endangered Languages spoken in Europe	18,700
Carpathian-German	Additional Languages spoken in Europe	4,690
Mocheno	Endangered Languages spoken in Europe	1,900
Meskhetian	Additional Languages spoken in Europe	200

Table 2: EU languages not found in LREC, ACL, EMNLP and CL documents (2000-2020)

guages suffering from an extreme HLT research inequality including theoretically non endangered languages with a relevant number of speakers.

Table 3 included in the Appendix shows the EU languages ordered in decreasing number of the total sum of LREC, ACL, EMNLP and CL papers between 2000-2020 mentioning each language. Interestingly, all the three venues and the journal publish research papers that mention many different languages in quite similar distributions. Both tables 2 and 3 also show the classification given to each language in the ELE project regarding if they are Official EU Languages, Additional Languages spoken in Europe or Endangered Languages spoken in Europe. In the second and third of these groups, Additional or Endangered Languages, we can find official languages of non EU Member States like Norwegian or Turkish, co-official languages of European Regions like Frisian (Additional) or Scottish Gaelic (Endangered), languages with certain recognition in their respective regions despite not being co-official like Venetian (Additional) or Breton (Endangered), and languages with no official status or recognition at all like Sicilian (Additional) or Lombard (Endagered). It is also worth noting the presence of Catalan and Basque, co-official languages in their respective regions in the top levels of the list overtaking several Official EU languages with a bigger number of speakers. Also, Turkish as the highest ranking non EU State Official language, precedes several Official EU Languages but in this case with a remarkably higher number estimated speakers than them. Picard, Breton and Tatar, with 700,000, 206,000 and 20,550 estimated speakers respectively, are the topmost mentioned Endangered Languages in LREC, ACL, EMNLP and CL documents 2000-2020, way above of much more spoken *Aditional Languages* like Sicilian, Lombard or Venetian with 4.7 million, 3.9 million and 3.8 million estimated speakers respectively.

Figure 4 describes the evolution of the number of papers mentioning the 20 most mentioned EU languages per year in the 2000 to 2020 period, i.e., the *absolute metric*. We can observe an overall nice and relatively parallel evolution of the number of research papers mentioning each EU language, particularly in the case of the most spoken languages. From this figure we could conclude that, with the exception of English probably due to its global *lingua franca* nature, the bigger the number of European citizens living in a country where the language is official, the higher the position of the language in this characterisation HLT research equality. As expected, this *absolute* top 20 list includes some of the most spoken Official EU Languages, but also Turkish and Norwegian, languages with non official status in the EU, and Catalan and Basque, both of them *Additional Languages* spoken in Europe that enjoy full official status in their respective regions.

Figure 5 describes the evolution of the number of papers mentioning the top 20 EU languages mentioned on documents per million of estimated speakers, i.e., the *relative metric*. This *relative* top 20 list includes, as we could expect, mainly languages with lower number of speakers, some of them Official EU Languages like Estonian, Maltese, Irish, Czech, Danish, Latvian, Finnish and Slovene, and all of the rest are languages enjoying a certain degree of official status or recognition in their respective regions of reference. Also remarkably we can observe that Czech, Swedish, Norwegian, Finnish, Danish, Basque, Portuguese and Turkish are in both in the *absolute* and the *relative* top 20 language list, Basque being the only non-national Official EU Language one. It seems that the group of languages ranking in similarly high positions in both the *absolute* and the *relative metric*, also exhibits some sort of equality in HLT research among them.

Stepping a bit deeper in this *relative metric*, Figure 6 depicts the evolution of the number of research papers mentioning EU languages per million of speakers for the most spoken EU languages (over 10 million speakers in Europe) between 2000 and 2020. In this figure we can observe how languages with a lower number of estimated speakers rank consistently better than those languages with a higher number of estimated speakers. Taking English as a reference we can observe two different groups within these strongest languages. On

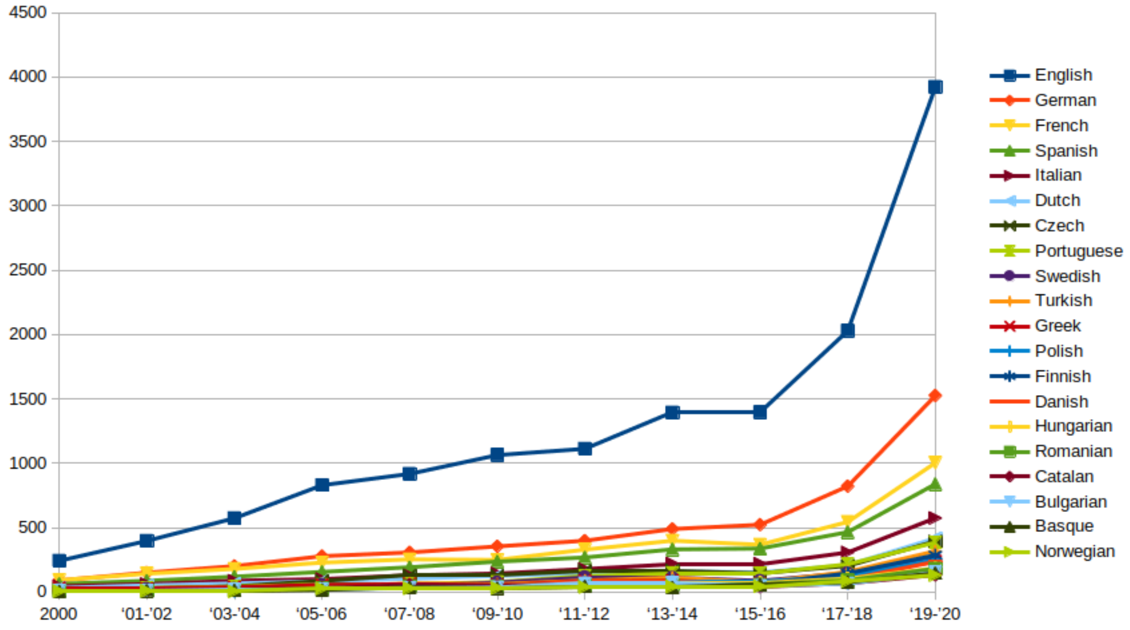

Figure 4: Evolution of mentions of European languages in LREC, ACL, EMNLP and CL documents 2000-2020.

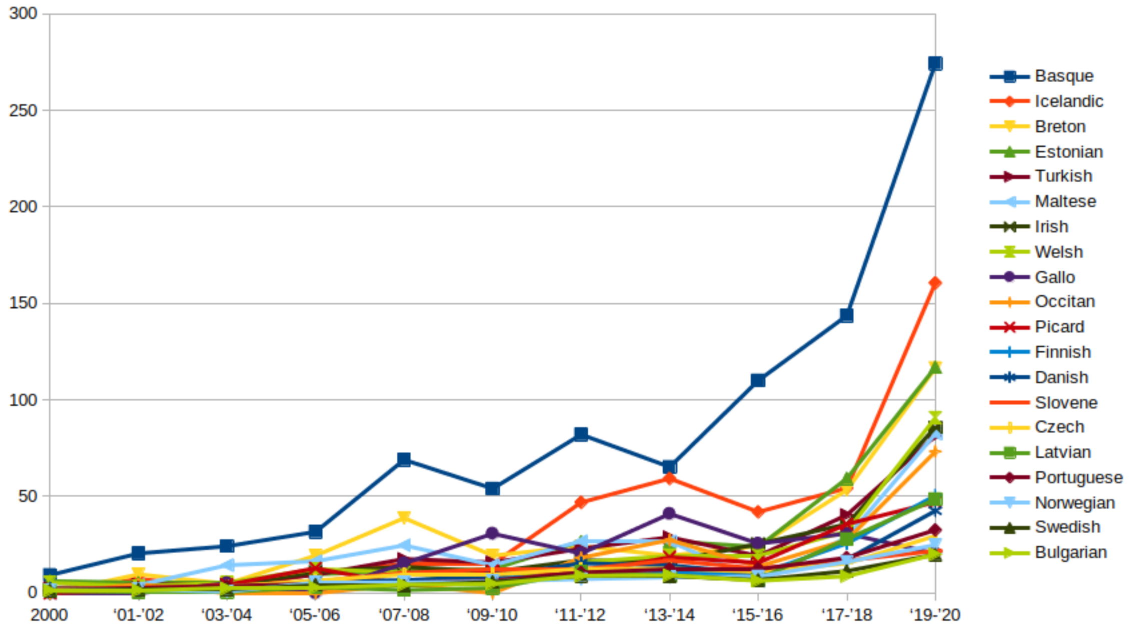

Figure 5: Evolution of mentions of European languages in LREC, ACL, EMNLP and CL documents 2000-2020 per million speakers.

one hand the ones on higher positions than English with Portuguese, Czech, Swedish, Greek, Dutch and Hungarian in this group, and those on lower positions than English with Spanish, German, Italian, Romanian, French, Serbian and Polish in this group. The existence of these two groups according to this metric may suggest the existence of a new inequality in HLT research in this case compared to the international *lingua franca*. Some strong European languages may be underrepresented and lagging too much behind English in HLT research in proportion to their demographic relevance in Europe.

6. Conclusions

This work proposes two quantitative metrics for measuring the HLT research equality of European lan-

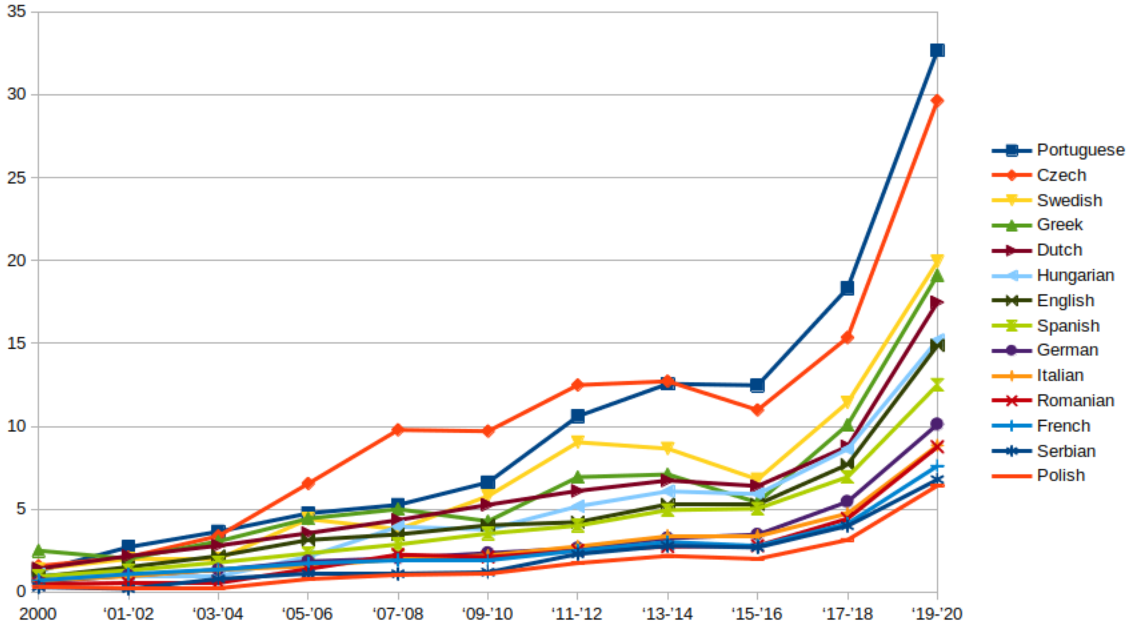

Figure 6: LREC, ACL, EMNLP and CL documents 2000-2020 mentioning the top EU languages (over 10 million speakers in the EU) per million speakers.

guages: an *absolute metric* counting the number of HLT scientific papers mentioning each language, and a *relative metric* counting the number of papers mentioning each language per million of European speakers. These two metrics do not pretend to measure the performance or effectiveness of the overall HLT research among languages.

The data gathered and analysed in this work suggests that despite the effort towards language equality of HLT research in Europe, there is still a large room for improvement. In fact, according to the proposed metrics on the selected data sources the European languages are largely unequal in HLT research. Nevertheless we have identified three groups of EU languages with a relatively homogeneous behaviour in terms of HLT research according to the proposed metrics. Each group comprises languages of quite a varying number of speakers: 1) a group of EU languages that we may describe equal in the vulnerability regarding HLT research ranking poorly in both the *absolute* and the *relative* metrics, in addition to the languages with no mention found. This group includes a long list of languages, some of them with a large number of speakers like Sicilian, Sardinian, Venetian, Alsatian Lombard or Romani; 2) a group of languages that appear in top positions in both proposed metrics and with Czech, Swedish, Norwegian, Finnish, Danish, Basque, Portuguese and Turkish as some clear representatives, and 3) a group of strong official languages with a large base of speakers ranking in a high position in the *absolute metric* and in an intermediate position in the *relative metric*, including among others German, French, Italian, Spanish and Dutch, that could be lagging behind English for its outstanding position in the *absolute metric*. There are of course also languages in intermediate positions between these groups.

As expected, we have observed that the combination of officialdom and a relevant number of speakers are positive conditions for a higher presence in HLT research. Also, not being a recognized language, at least regionally, burdens definitely its equality with respect to the ones that enjoy some degree of officialdom, no matter the size of the population speaking that language. On the other hand, it seems that regionally recognised languages can perform as good as national Official EU Languages.

Finally, we can conclude that the combination of both indicators can be of utility for measuring the HLT research equality.

Next, we plan to set up a dashboard web site to interact and order the data by its different parameters. Additionally, we plan to perform an in-depth analysis of the sources of inequalities for a better future support and understanding of the HLT research equality in Europe and other multilingual regions in the world.[13]

7. Bibliographical References

Ahmed, N. and Wahed, M. (2020). The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Blasi, D., Anastasopoulos, A., and Neubig, G. (2021). Systematic inequalities in language technology per-

[13] The data and code will be released upon acceptance.

formance across the world's languages. *arXiv e-prints*, pages arXiv–2110.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press, Cambridge, MA, USA.

Joscelyne, A. and Lockwood, R. (2003). *Benchmarking HLT progress in Europe*. EUROMAP Language Technologies, Center for Sprogteknologi.

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.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Moseley, C. (2010). Atlas of the world's languages in danger, 3rd edn.

Pastor, R. (2018). *Language equality in the digital age : towards a human language project*. European Parliament.

Ponti, E. M., O'Horan, H., Berzak, Y., Vulić, I., Reichart, R., Poibeau, T., Shutova, E., and Korhonen, A. (2019). Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601, 09.

Rehm, G. and Hegele, S. (2018). Language technology for multilingual Europe: An analysis of a large-scale survey regarding challenges, demands, gaps and needs. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. 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.

Rehm, G., Uszkoreit, H., Dagan, I., Goetcherian, V., Dogan, M. U., Mermer, C., Váradi, T., Kirchmeier-Andersen, S., Stickel, G., Jones, M. P., Oeter, S., and Gramstad, S. (2014). An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In Laurette Pretorius, et al., editors, *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL 2014)*, pages 30–37, Reykjavik, Iceland.

Rehm, G., Marheinecke, K., Hegele, S., Piperidis, S., Bontcheva, K., Hajič, J., Choukri, K., Vasiļjevs, A., Backfried, G., Prinz, C., Gómez-Pérez, J. M., Meertens, L., Lukowicz, P., van Genabith, J., Lösch, A., Slusallek, P., Irgens, M., Gatellier, P., Köhler, J., Le Bars, L., Anastasiou, D., Auksoriūtė, A., Bel, N., Branco, A., Budin, G., Daelemans, W., De Smedt, K., 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., Skadiņa, I., Tadić, M., Tufiș, D., Váradi, T., Vider, K., Way, A., and Yvon, F. (2020). The European language technology landscape in 2020: Language-centric and human-centric AI for cross-cultural communication in multilingual Europe. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3322–3332, Marseille, France. European Language Resources Association.

Sayers, D., Sousa-Silva, R., Höhn, S., Ahmedi, L., Allkivi-Metsoja, K., Anastasiou, D., Beňuš, Štefan; Bowker, L., Bytyçi, E., Catala, A., Çepani, A., Chacón-Beltrán, Rubén; Dadi, S., Dalipi, F., Despotovic, V., Doczekalska, A., Drude, S., Fort, Karën; Fuchs, R., Galinski, C., Galinski, C., Galinski, C., Gobbo, F., Gungor, T., Guo, S., Höckner, K., Láncos, P., Libal, T., Jantunen, T., Jones, D., Klimova, B., Korkmaz, E., Maučec, M. S., Melo, M., Meunier, F., Migge, B., Mititelu, V. B., Névéol, Aurélie; Rossi, A., Pareja-Lora, A., Sanchez-Stockhammer, C.; Şahin, A., Soltan, A., Soria, C., Shaikh, S., Turchi, M., Yildirim Yayilgan, S., Bessa, M., Cabral, L., Coler, M., Liebeskind, C., Kernerman, I., Rousi, R., and Prys, C. (2021). The dawn of the human-machine era : A forecast of new and emerging language technologies. Technical report, LITHME project.

Torfi, A., Shirvani, R. A., Keneshloo, Y., Tavvaf, N., and Fox, E. A. (2020). Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Appendix

Language	**Classification**	**Speakers**	**LREC**	**ACL**	**EMNLP**	**CL**	**Total**
English	Official European Union Languages	263,835,370	4,676	4,839	3,837	531	13,883
German	Official European Union Languages	150,888,580	2,013	1,602	1,304	227	5,146
French	Official European Union Languages	131,992,030	1,783	1,027	803	182	3,795
Spanish	Official European Union Languages	67,144,190	1,377	872	723	131	3,103
Italian	Official European Union Languages	65,019,690	1,004	554	429	87	2,074
Dutch	Official European Union Languages	23,918,840	737	423	310	86	1,556
Czech	Official European Union Languages	13,295,420	593	510	361	55	1,519
Portuguese	Official European Union Languages	11,787,500	627	358	269	53	1,307
Swedish	Official European Union Languages	12,947,670	449	267	209	49	974
Turkish	Additional Languages spoken in Europe	3,905,040	302	342	261	62	967
Greek	Official European Union Languages	12,399,170	391	221	206	49	867
Polish	Official European Union Languages	39,415,080	353	220	153	32	758
Finnish	Official European Union Languages	5,682,630	263	267	183	32	745
Danish	Official European Union Languages	5,563,120	252	234	213	19	718
Hungarian	Official European Union Languages	12,177,260	254	219	155	28	656
Romanian	Official European Union Languages	20,776,510	265	194	114	21	594
Catalan	Additional Languages spoken in Europe	8,973,480	274	128	117	29	548
Bulgarian	Official European Union Languages	7,570,230	212	173	122	26	533
Basque	Additional Languages spoken in Europe	536,000	191	130	133	20	474
Norwegian	Additional Languages spoken in Europe	5,254,060	208	121	102	21	452
Estonian	Official European Union Languages	1,128,990	146	104	80	13	343
Croatian	Official European Union Languages	6,590,290	160	84	64	9	317
Irish	Official European Union Languages	1,176,730	102	86	67	7	262
Slovene	Official European Union Languages	2,195,790	118	79	52	10	259
Slovak	Official European Union Languages	7,174,580	115	63	58	5	241
Serbian	Additional Languages spoken in Europe	10,025,456	112	55	61	5	233
Latvian	Official European Union Languages	1,933,100	98	64	47	9	218
Lithuanian	Official European Union Languages	2,793,100	70	76	36	3	185
Icelandic	Additional Languages spoken in Europe	404,683	85	57	20	5	167
Galician	Additional Languages spoken in Europe	2,335,000	80	45	28	2	155
Welsh	Additional Languages spoken in Europe	562,000	49	37	29	9	124
Maltese	Official European Union Languages	485,110	66	37	13	3	119
Picard	Endangered Languages spoken in Europe	700,000	36	39	35	3	113
Macedonian	Additional Languages spoken in Europe	1,553,203	40	30	16	5	91
Breton	Endangered Languages spoken in Europe	206,000	32	18	15	3	68
Tatar	Endangered Languages spoken in Europe	20,550	17	14	18	1	50
Faroese	Additional Languages spoken in Europe	76,587	23	13	13	0	49
Frisian	Additional Languages spoken in Europe	883,000	22	22	3	1	48
Sorbian	Endangered Languages spoken in Europe	19,970	16	6	24	1	47
Asturian	Endangered Languages spoken in Europe	560,000	21	13	4	0	38
Occitan	Additional Languages spoken in Europe	218,310	25	7	5	0	37
Gallo	Endangered Languages spoken in Europe	195,000	10	12	12	3	37
Romani	Endangered Languages spoken in Europe	3,755,600	14	15	7	0	36
Yiddish	Endangered Languages spoken in Europe	10,977	13	14	3	2	32
Lombard	Endangered Languages spoken in Europe	3,903,000	22	5	3	0	30
Luxembourgish	Additional Languages spoken in Europe	510,900	15	9	4	0	28
Cornish	Endangered Languages spoken in Europe	600	6	13	5	3	27
Scottish Gaelic	Endangered Languages spoken in Europe	57,400	12	4	9	1	26
Venetian	Additional Languages spoken in Europe	3,850,000	13	6	1	0	20
Aragonese	Endangered Languages spoken in Europe	30,000	8	6	3	0	17
Sardinian	Endangered Languages spoken in Europe	1,200,000	10	4	2	1	17
Ladin	Endangered Languages spoken in Europe	31,000	8	6	1	0	15
Sicilian	Additional Languages spoken in Europe	4,700,000	8	4	3	0	15
Karelian	Endangered Languages spoken in Europe	5,000	7	4	3	0	14
Saami	Endangered Languages spoken in Europe	22,430	7	3	4	0	14
Manx	Endangered Languages spoken in Europe	1,660	5	4	2	1	12
Alsatian	Additional Languages spoken in Europe	600,000	8	0	2	1	11

Table 3: Number of LREC, ACL, EMNLP and CL documents 2000-2020 mentioning EU languages (languages with over 10 documents mentioning them)