Measuring HLT Research Equality of European Languages

Anonymous ACL submission

Abstract

This work explores quantitatively the equality of the languages of the European Union in the field of HLT. Our ultimate goal is to investigate European language diversity and identify low-resource and endangered languages taking into account the research papers of the main HLT conferences. This framework has been selected with the goal to identify potential inequalities among theoretically similarly capable languages in terms of available social and economical resources as well as political status. We have identified several groups of EU languages in terms of HLT research equality, each group comprising languages of very varying number of speakers. We have discovered a relative equality among surprisingly different languages in terms of speaker base and also relevant inequalities within the most spoken languages. All data and code will be released upon acceptance.

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, including languages of relevant trade partners and immigrant communities. The fact that several of the regional languages enjoy the same level of officialdom in their respective regions as the corresponding EU Member State language, 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 a relevant extent of the rest of non-official regional or minority languages, are both European particularities not easily found in other societies in the world. One of the reasons for these diversity and public support is that multilingualism is one of the core values of an 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 (TFUE), 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.

This multilingual nature of the EU is considered to be one of the union's differentiating elements and a key competitive advantage, but the singularity of European multilingualism comes at the extent on which 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 matter gets 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 inter-dependence among them, but each one with a different profile and mix of coexisting languages.

The language equality is a vibrant and remarkable challenge, and a research field that is building it's own foundations. This work intends to contribute to both the challenge and the emerging research field through the deliberation about the equality of European languages in their digital facet, particularly in the field of Human Language Technologies (HLT).

In recent years, the HLT community has developed 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 implement 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 firms 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 analyses of the HLT research equality of the European languages. Finally, Section 6 summarizes our main findings and presents our future work.

2 Related work

Given the role of LT in everyone's daily lives, many LT practitioners are directly concerned by language diversity in LT research and development.[1] For instance, (Sayers et al., 2021) emphasise a range of groups who will be disadvantaged and issues of inequality. Important issues of security and privacy will accompany new LT. Looking ahead, they see many intriguing opportunities and new capabilities, but a range of other 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 LT 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 LT research (Ponti et al., 2019). (Blasi et al., 2021) study the systematic inequalities in LT across World languages. After English, a handful of Western European Languages dominate the field -in particular German, French and Spanish- 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 LT.

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 LT (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., 2020a) present an overview of various European LT and AI reports. Finally, (Rehm et al., 2020b) perform an extensive qualitative analysis of the landscape of research on Language Technologies in all the Member countries of the EU.

Our work intends to explore quantitatively the equality levels within languages of the EU, complementing the latter work, and with the goal to unveil potential inequalities among theoretically top performing languages that would be classified in the same tier comparing the to the whole universe of languages in the world.

3 Initial hypothesis

HLT are, themselves, regarded as language agnostic or inherently equal to any language. This field of knowledge is not particularly dependant on relevant capital investments, availability of natural resources or geopolitical factors. Research, development and innovation in HLT is, generally, affordable and equally accessible for societies that have reached certain level of human and economic

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

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 offers to the variety of European languages creates a unique case of theoretical equality among these different languages.

The initial hypothesis of his work is that, particularly in the field of HLT, the languages of the EU should show a relevant degree of equality, at least within the languages with the same level of official support, 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 collectives (policy makers, academy, industry and any other) that could have legitimate interest in correcting the divergence. Also, on the other hand, could confirm the effectiveness of existing scientific, regulatory, policy and societal dynamics in the purpose of achieving the language equality.

Finally, the decided focus on HLT for the study is expected to be further beneficial contributing to the 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 the equality with others, may converge faster to equality taking advantage of HLT, but failing or performing poorly on HLT may be an unbridgeable barrier to gain overall language equality, or even a menace to loose feet in the future and plummet, even for currently well-resourced languages that could not perform too well in this subject.

4 Selected Languages and Measurement Method

The selected languages for the study are those identified as languages of the EU in the European Language Equality Project (ELE).[2] 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. Figure 1 describes, sorted by estimations of global number of total speakers, the list of languages of the EU considered in ELE project and the breakdown of importance of each language considering only the global number of speakers. The estimations of number of speakers have been obtained from the online encyclopedia of writing systems and languages Omniglot,[3] and open searches on the internet for languages not included in this database. More than 90% of speakers are concentrated in 8 languages out of 67 main EU languages. This top group includes 4 global languages, English, Spanish, French and Portuguese, languages born in Europe but with more speakers abroad than in their countries of origin. Also, within the top 8 languages, we can observe a steep gradient being almost half of them English speakers and approximately 2% of them speakers of Polish. Considering only this metric, languages of the EU are inherently and deeply non equal.

The basic indicator we have selected to measure the equality among languages in the field of HLT is the number of scientific documents that mention each language published in the period from 2000 to 2020. 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)[4], the Annual Meeting of the Association for Computational Linguistics (ACL)[5], the Conference on Empirical Methods in Natural Language Processing (EMNLP)[6], and the Computational Linguistics Journal (CL)[7]. We have crawled all documents published in these venues from 2000 to 2020 available in the ACL Anthology website[8], extracted the text of these files, and have found what EU languages are mentioned in each document, according the list developed by the ELE project and after filtering proper nouns that are the same as EU languages but not refer to a Language, e.g. "Basque" in the name "University of the Basque Country" does not count as mention of Basque language. Table 1 shows the number of research papers processed from each source.

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

[3] https://omniglot.com/

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

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

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

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

[8] https://aclanthology.org/

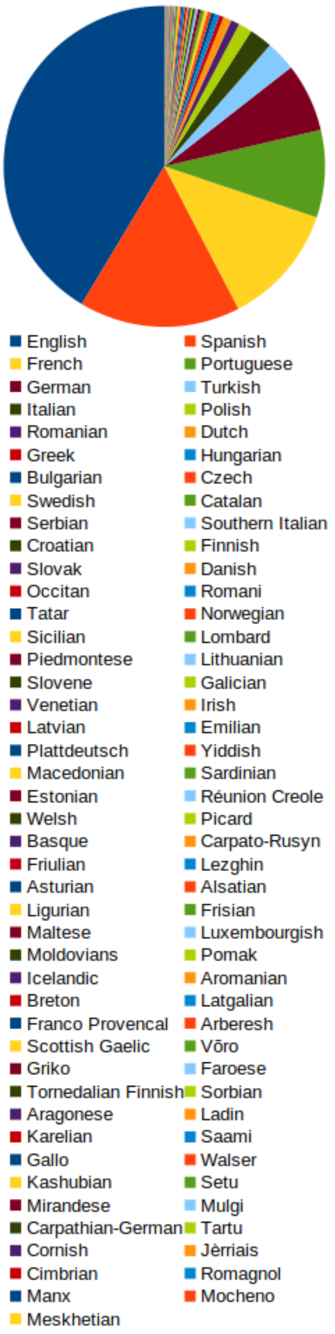

Figure 1: Speakers per language of the EU.

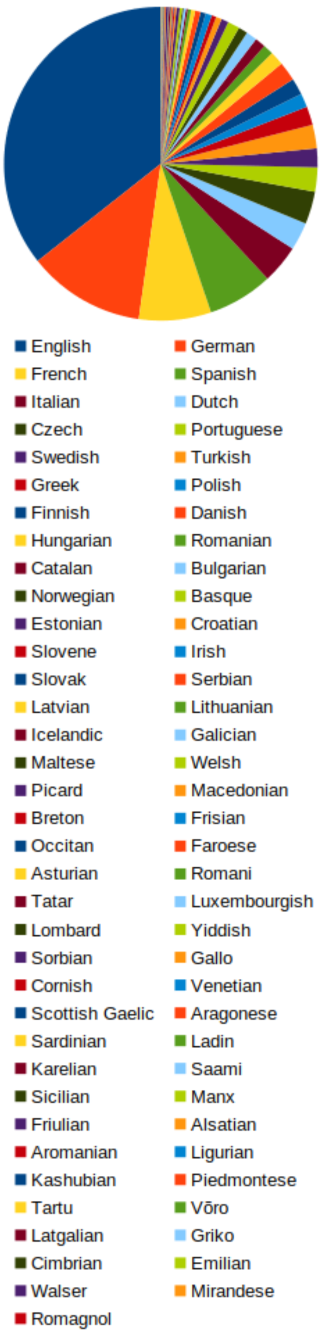

Figure 2: Documents mentioning languages of the EU (only languages with published documents).

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

Figure 2 shows the breakdown of European languages sorted by total number of documents mentioning each language. This figure shows intuitively a slightly lower degree of inequality compared to the one depicted in Figure 1, but the inherent inequality among languages still remains clear. It is also worth noting that in this characterisation German and French, despite having a lower number of speaker, have more documents that mention them than Spanish, positioned as fourth most men-

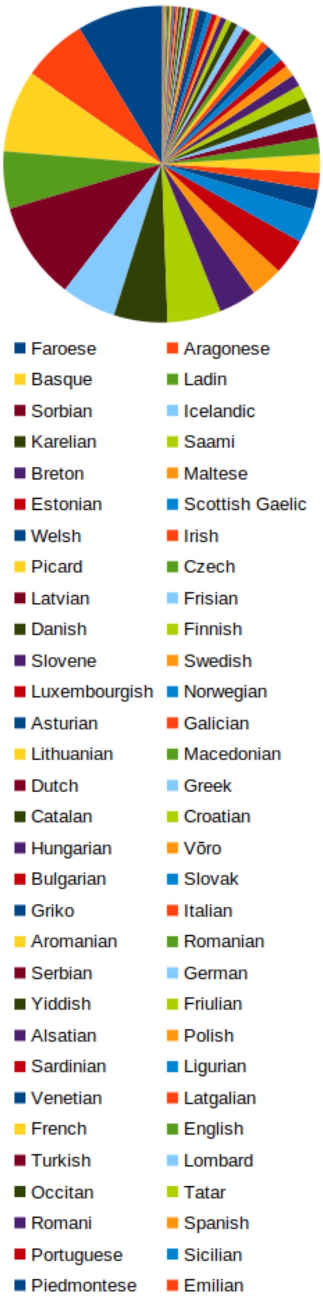

Figure 3: Documents mentioning languages of the EU per million of speakers (only languages with published documents and with over 30.000 speakers).

tioned EU language in these sources. Also Italian, Dutch and Czech perform better than more spoken languages like Portuguese and Turkish. These variations in the relative position of each language in these rankings advance that there could be inequalities of different nature among languages, not affecting only low-resource and endangered languages but also some of the strongest and most spoken languages in the world.

To further distill the analysis, we remove from the characterisation the natural inequality of languages coming from their number of speakers. Figure 3 shows the breakdown of number of documents mentioning each language per million of speakers of that language. We have removed from this ranking languages below 30.000 speakers because these low numbers in the denominator of the ratio introduce too noisy and non representative distortions in the comparison with languages with several millions of speakers in the denominator. Qualitatively observing the pie chart, and comparing it to the ones in figures 1 and 2, we can conclude that the differences between languages in this characterisation, eliminated the bias of the number of speakers, are lower showing a higher equality levels among EU languages overall. 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, languages with lower numbers of speakers like Aragonese, Faroese or Basque rise to the top of the list. Also in this case, we can observe different behaviours among languages. In the previous 2 we have observed inequalities among the strongest languages, and now we can observe different dynamics and performances also within the group of less spoken, potentially endangered languages. We can observe some of these languages performing in the top positions, and also some of them in the lowest positions showing also inequalities among small languages.

5 Analysis of language equality

Table 2 includes the EU languages identified in ELE project for which no mentions have been found in the sources. Table 3 included in the Appendix A 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. Both tables 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 groups we can find official languages of non EU 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 officialdom or recognition at all like Sicilian (Additional) or Lombard (Endagered).

In Table 2 we find sixteen languages classified by ELE project as Additional Languages and Endangered Languages spoken in Europe, with the remarkable presence of Southern Italian and Plattdeutsch, with 7,500,000 and 1,700,000 estimated speakers respectively. Less spoken but still relevant languages like Carpato-Rusyn, Lezghin and Réunion Creole, all of them above 600,000 estimated speakers are also included in this list. The existence of this list brings to surface the first and most relevant tier of non equality in the group of EU languages in the field of HLT: the ones not even mentioned once in the most relevant HLT conferences in the world in the last 20 years of scientific research. Also note that, contrary to what could be expected, most of the languages included in this list are not considered endangered. None of the languages in this list has a officially recognised status by the national or regional governments of the areas where they are spoken.

Following the analysis regarding the level of officialdom of languages, in Table 3, it is also worth noting the presence of the Catalan and Basque regional co-official languages 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 5,200,000 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, Piedmontese or Emilian with 5 million, 3 million and 1,7 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. We can observe an overall nice and relatively parallel evolution of the number of papers mentioning each EU language, particularly in the case of the most spoken languages. Anyhow, this graph shows that the gap between languages in this measurement tends to grow in time. This scenario depicts an evolution on which the inequality between EU languages in the field HLT tends to increase in time, favouring those languages that are particularly strong in the field like English, German and French, versus weaker ones like Spanish, Italian and Portuguese. Also, 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 better the performance of the language in this characterisation. This "absolute" top 20 list includes some of the most spoken Official EU Languages, as we could intuitively expect, but also Turkish and Norwegian, languages with non officialdom in the EU, and Catalan and Basque, both of them *Additional Languages* spoken in Europe that enjoy full officialdom in their respective regions.

In the Section 4 we have concluded that for making the most non-biased analysis possible, the comparison between languages should be based on measurements relative to the number of speaker of each language. Figure 5 describes the evolution of the number of papers mentioning the top 20 EU languages on documents mentioning them per million of estimated speakers. This "relative" top 20 list includes, as we could expect, mainly languages with lower number of speakers, some of them Official EU Languages like Icelandic, Estonian, Maltese, Irish, Czech, Danish, Latvian, Finnish ans Slovene, and all of the rest are languages enjoying a certain degree of officialdom or recognition in their respective regions of reference. Also remarkably we can observe that Dutch, Czech, Finnish, Danish and Basque are in both in the "absolute" and the "relative" top 20 language list, being Basque the only not national Official EU Language but only regionally official in the Basque Country.

Stepping a bit deeper in this subject, Figure 6 depicts the evolution of the number of research papers mentioning EU languages per million of speakers for the 10 most spoken EU languages between 2000 and 2020, a.i., the apparently less biased way to measure the equality among languages. In this figure we can observe how languages with a lower number of estimated speakers outperform consistently 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 one hand the ones performing better than English with Dutch, Italian, German, Romanian, Polish and Turkish in this group, and those performing worst than English with French, Spanish and Portuguese in this group.

Language	ELE Classification	Speakers
Southern Italian	Additional Languages spoken in Europe	7,500,000
Plattdeutsch	Additional Languages spoken in Europe	1,700,000
Réunion Creole	Additional Languages spoken in Europe	800,000
Carpato-Rusyn	Additional Languages spoken in Europe	636,000
Lezghin	Additional Languages spoken in Europe	600,000
Moldovians	Additional Languages spoken in Europe	400,000
Pomak	Additional Languages spoken in Europe	351,000
Franco Provencal	Endangered Languages spoken in Europe	140,000
Arberesh	Endangered Languages spoken in Europe	100,000
Tornedalian Finnish	Additional Languages spoken in Europe	60,000
Setu	Endangered Languages spoken in Europe	12,500
Mulgi	Additional Languages spoken in Europe	10,000
Carpathian-German	Additional Languages spoken in Europe	5,500
Jèrriais	Endangered Languages spoken in Europe	2,700
Mocheno	Endangered Languages spoken in Europe	1,700
Meskhetian	Additional Languages spoken in Europe	500

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

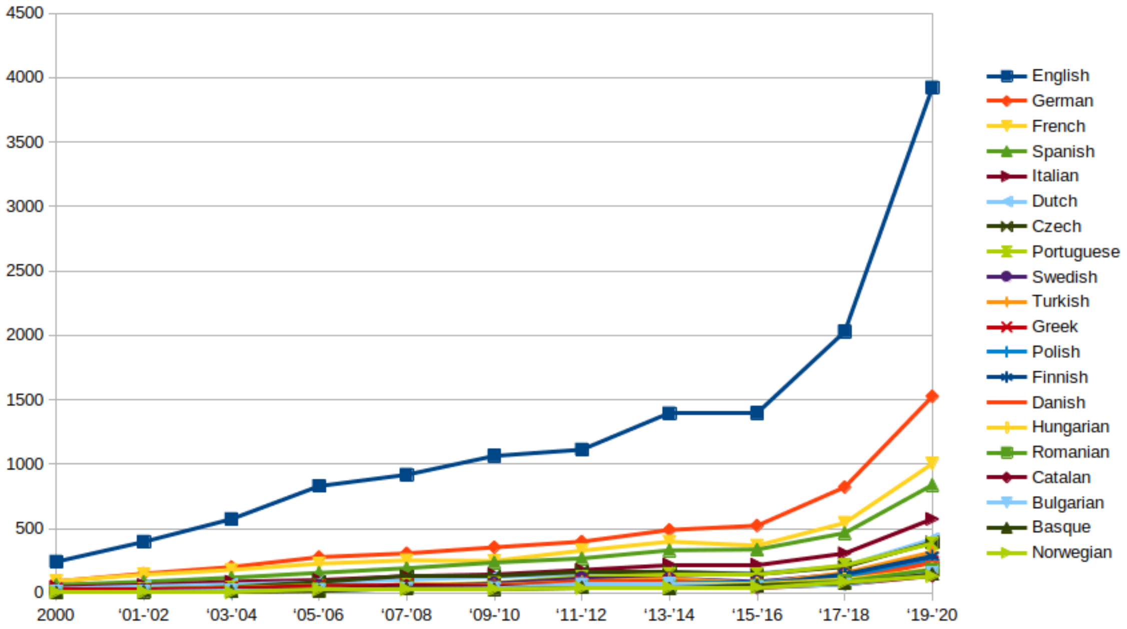

Figure 4: Evolution of mentions of European languages in LREC, ACL, EMNLP and CL documents 2000-2020.

The latter happen to be the EU languages with most non European speakers in the world, and this seems to negatively affect these languages in this comparison.

6 Conclusions

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.[9] We have identified several tiers of EU languages in terms of equality on HLT, each group comprising languages of very varying number of speakers: 1) the most endangered ones not being mentioned even once in the HLT research papers, having in common that none of them enjoys any level of officialdom, 2) strong languages weakly performing in the field relatively to their number of speakers, having in common a strong base of speakers outside Europe, and 3) relatively equal languages. As expected, we have observed that the combination of officialdom and a relevant population speaking a particular language in Europe are positive conditions. Also, not being a recognized language, at least a regionally, burdens definitely its inequality with respect the ones that enjoy some degree of officialdom. No matter the size of the population speaking that language. On the other hand, regionally recognised languages

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

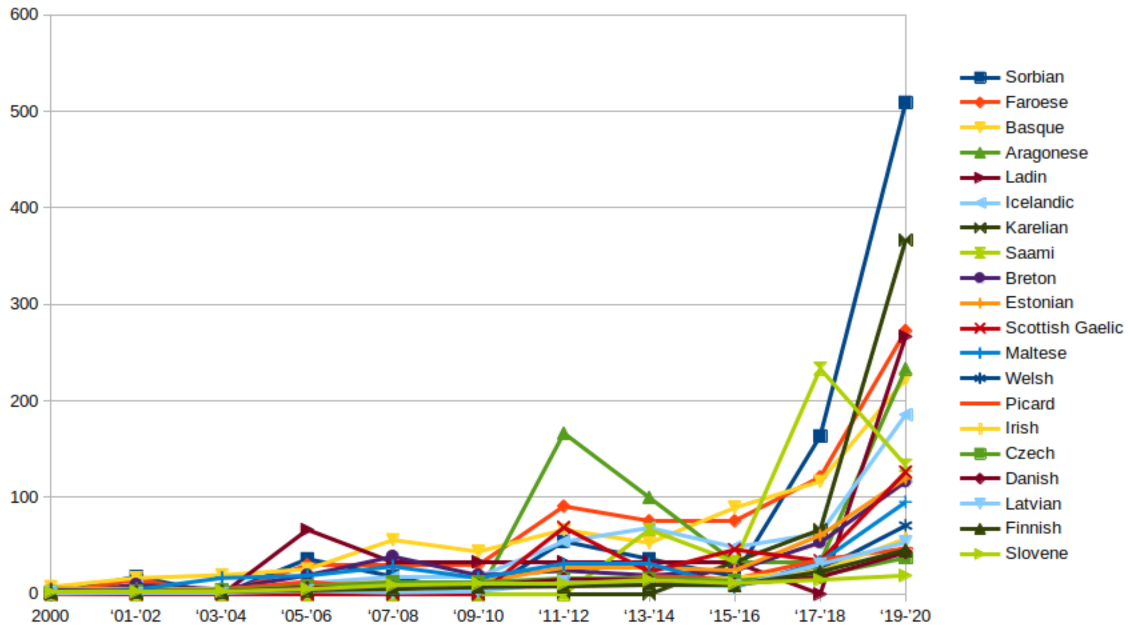

Figure 5: Evolution of mentions of European languages in LREC, ACL, EMNLP and CL documents 2000-2020 per million speakers.

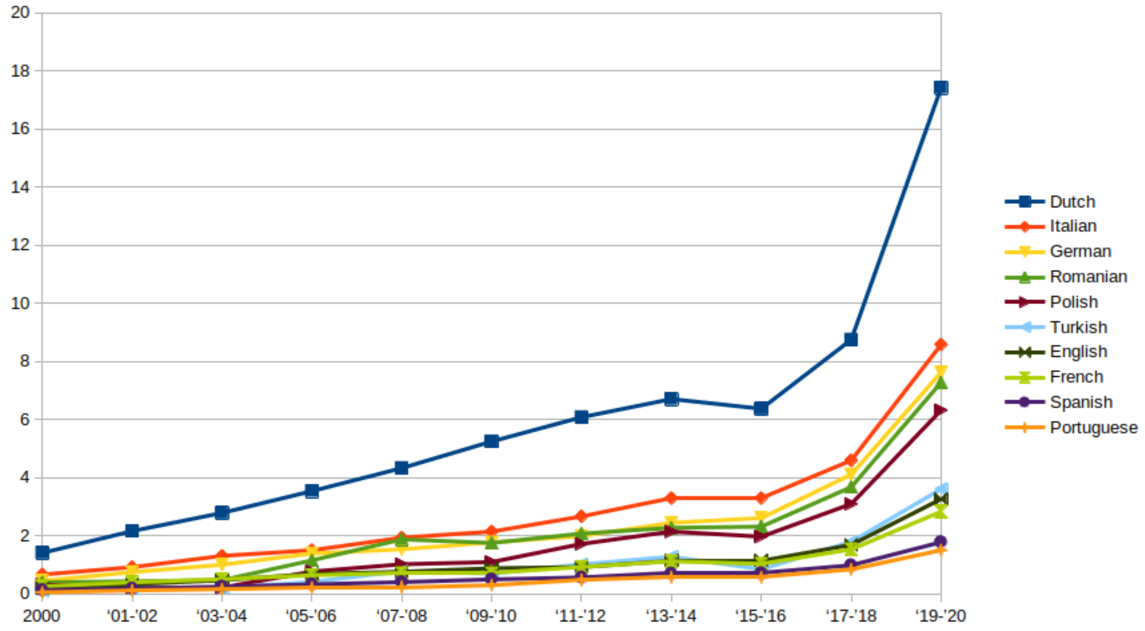

Figure 6: LREC, ACL, EMNLP and CL documents 2000-2020 mentioning the 10 most spoken EU languages per million speakers.

can perform as good as national Official EU Languages. 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 these remarkable equities and inequalities for a better future support and understanding of the language equality in HLT in Europe and other multilingual regions in the world.

References

Nur Ahmed and Muntasir Wahed. 2020. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

Damián Blasi, Antonios Anastasopoulos, and Graham Neubig. 2021. Systematic inequalities in language technology performance across the world's languages. *arXiv e-prints*, pages arXiv–2110.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Ian J. Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press, Cambridge, MA, USA.

Andrew Joscelyne and Rose Lockwood. 2003. *Benchmarking HLT progress in Europe*. EUROMAP Language Technologies, Center for Sprogteknologi.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. *Computational Linguistics*, 45(3):559–601.

Georg Rehm and Stefanie Hegele. 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, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajič, Khalid Choukri, Andrejs Vasiljevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez-Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadiņa, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. 2020a. 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.

Georg Rehm, Katrin Marheinecke, Stefanie Hegele, Stelios Piperidis, Kalina Bontcheva, Jan Hajič, Khalid Choukri, Andrejs Vasiljevs, Gerhard Backfried, Christoph Prinz, José Manuel Gómez-Pérez, Luc Meertens, Paul Lukowicz, Josef van Genabith, Andrea Lösch, Philipp Slusallek, Morten Irgens, Patrick Gatellier, Joachim Köhler, Laure Le Bars, Dimitra Anastasiou, Albina Auksoriūtė, Núria Bel, António Branco, Gerhard Budin, Walter Daelemans, Koenraad De Smedt, Radovan Garabík, Maria Gavriilidou, Dagmar Gromann, Svetla Koeva, Simon Krek, Cvetana Krstev, Krister Lindén, Bernardo Magnini, Jan Odijk, Maciej Ogrodniczuk, Eiríkur Rögnvaldsson, Mike Rosner, Bolette Pedersen, Inguna Skadiņa, Marko Tadić, Dan Tufiș, Tamás Váradi, Kadri Vider, Andy Way, and François Yvon. 2020b. 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.

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

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Váradi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, and Sigve Gramstad. 2014. An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age". In *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.

Dave Sayers, Rui Sousa-Silva, Sviatlana Höhn, Lule Ahmedi, Kais Allkivi-Metsoja, Dimitra Anastasiou, Lynne Beňuš, Štefan; Bowker, Eliot Bytyçi, Alejandro Catala, Anila Çepani, Sami Chacón-Beltrán, Rubén; Dadi, Fisnik Dalipi, Vladimir Despotovic, Agnieszka Doczekalska, Sebastian Drude, Robert Fort, Karën; Fuchs, Christian Galinski, Christian Galinski, Christian Galinski, Federico Gobbo, Tunga Gungor, Siwen Guo, Klaus Höckner, PetraLea Láncos, Tomer Libal, Tommi Jantunen, Dewi

Jones, Blanka Klimova, EminErkan Korkmaz, Mirjam Sepesy Maučec, Miguel Melo, Fanny Meunier, Bettina Migge, Verginica Barbu Mititelu, Arianna Névéol, Aurélie; Rossi, Antonio Pareja-Lora, Aysel Sanchez-Stockhammer, C.; Şahin, Angela Soltan, Claudia Soria, Sarang Shaikh, Marco Turchi, Sule Yildirim Yayilgan, Maximino Bessa, Luciana Cabral, Matt Coler, Chaya Liebeskind, Ilan Kernerman, Rebekah Rousi, and Cynog Prys. 2021. The dawn of the human-machine era : A forecast of new and emerging language technologies. Technical report, LITHME project.

Amirsina Torfi, Rouzbeh A Shirvani, Yaser Keneshloo, Nader Tavvaf, and Edward A Fox. 2020. Natural language processing advancements by deep learning: A survey. *arXiv preprint arXiv:2003.01200*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

A Appendix

Language	**Classification**	**Speakers**	**LREC**	**ACL**	**EMNLP**	**CL**	**Total**
English	Official EU Languages	1,200,000,000	4,676	4,839	3,837	531	13,883
German	Official EU Languages	200,000,000	2,013	1,602	1,304	227	5,146
French	Official EU Languages	354,000,000	1,783	1,027	803	182	3,795
Spanish	Official EU Languages	470,000,000	1,377	872	723	131	3,103
Italian	Official EU Languages	67,000,000	1,004	554	429	87	2,074
Dutch	Official EU Languages	24,000,000	737	423	310	86	1,556
Czech	Official EU Languages	10,500,000	593	510	361	55	1,519
Portuguese	Official EU Languages	255,000,000	627	358	269	53	1,307
Swedish	Official EU Languages	10,000,000	449	267	209	49	974
Turkish	Additional Languages spoken in Europe	88,000,000	302	342	261	62	967
Greek	Official EU Languages	13,100,000	391	221	206	49	867
Polish	Official EU Languages	40,000,000	353	220	153	32	758
Finnish	Official EU Languages	6,300,000	263	267	183	32	745
Danish	Official EU Languages	5,500,000	252	234	213	19	718
Hungarian	Official EU Languages	13,000,000	254	219	155	28	656
Romanian	Official EU Languages	25,000,000	265	194	114	21	594
Catalan	Additional Languages spoken in Europe	9,500,000	274	128	117	29	548
Bulgarian	Official EU Languages	12,000,000	212	173	122	26	533
Basque	Additional Languages spoken in Europe	660,000	191	130	133	20	474
Norwegian	Additional Languages spoken in Europe	5,000,000	208	121	102	21	452
Estonian	Official EU Languages	1,100,000	146	104	80	13	343
Croatian	Official EU Languages	6,700,000	160	84	64	9	317
Irish	Official EU Languages	1,760,000	102	86	67	7	262
Slovene	Official EU Languages	2,500,000	118	79	52	10	259
Slovak	Official EU Languages	5,600,000	115	63	58	5	241
Serbian	Additional Languages spoken in Europe	9,500,000	112	55	61	5	233
Latvian	Official EU Languages	1,750,000	98	64	47	9	218
Lithuanian	Official EU Languages	2,900,000	70	76	36	3	185
Icelandic	Additional Languages spoken in Europe	350,000	85	57	20	5	167
Galician	Additional Languages spoken in Europe	2,400,000	80	45	28	2	155
Welsh	Additional Languages spoken in Europe	720,000	49	37	29	9	124
Maltese	Official EU Languages	420,000	66	37	13	3	119
Picard	Endangered Languages spoken in Europe	700,000	36	39	35	3	113
Macedonian	Additional Languages spoken in Europe	1,400,000	40	30	16	5	91
Breton	Endangered Languages spoken in Europe	206,000	32	18	15	3	68
Tatar	Endangered Languages spoken in Europe	5,200,000	17	14	18	1	50
Faroese	Additional Languages spoken in Europe	66,000	23	13	13	0	49
Frisian	Additional Languages spoken in Europe	470,000	22	22	3	1	48
Sorbian	Endangered Languages spoken in Europe	55,000	16	6	24	1	47
Asturian	Endangered Languages spoken in Europe	550,000	21	13	4	0	38
Occitan	Additional Languages spoken in Europe	5,500,000	25	7	5	0	37
Gallo	Endangered Languages spoken in Europe	28,000	10	12	12	3	37
Romani	Endangered Languages spoken in Europe	5,500,000	14	15	7	0	36
Yiddish	Endangered Languages spoken in Europe	1,500,000	13	14	3	2	32
Lombard	Endangered Languages spoken in Europe	3,900,000	22	5	3	0	30
Luxembourgish	Additional Languages spoken in Europe	400,000	15	9	4	0	28
Cornish	Endangered Languages spoken in Europe	3,000	6	13	5	3	27
Scottish Gaelic	Endangered Languages spoken in Europe	87,000	12	4	9	1	26
Venetian	Additional Languages spoken in Europe	2,000,000	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	30,000	8	6	1	0	15
Sicilian	Additional Languages spoken in Europe	5,000,000	8	4	3	0	15
Karelian	Endangered Languages spoken in Europe	30,000	7	4	3	0	14
Saami	Endangered Languages spoken in Europe	30,000	7	3	4	0	14
Manx	Endangered Languages spoken in Europe	1,800	5	4	2	1	12
Alsatian	Additional Languages spoken in Europe	548,000	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)