Systematic Inequalities in Language Technology Performance across the World's Languages

Damián Blasi
Harvard University
dblasi@fas.harvard.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Graham Neubig
Carnegie Mellon University
gneubig@cs.cmu.edu

Abstract

Natural language processing (NLP) systems have become a central technology in communication, education, medicine, artificial intelligence, and many other domains of research and development. While the performance of NLP methods has grown enormously over the last decade, this progress has been restricted to a minuscule subset of the world's ≈6,500 languages. We introduce a framework for estimating the global utility of language technologies as revealed in a comprehensive snapshot of recent publications in NLP. Our analyses involve the field at large, but also more in-depth studies on both user-facing technologies (machine translation, language understanding, question answering, text-to-speech synthesis) as well as foundational NLP tasks (dependency parsing, morphological inflection). In the process, we (1) quantify disparities in the current state of NLP research, (2) explore some of its associated societal and academic factors, and (3) produce tailored recommendations for evidence-based policy making aimed at promoting more global and equitable language technologies.[1]

1 Introduction

The past decade has seen a rapid advance in natural language processing (NLP); it has grown from a relatively technical niche to a fundamental tool in virtually all domains that involve language data in any shape or form. NLP is now instrumental to not only bread-and-butter applications such as translation and question answering, but also tasks as wide ranging as detection of neurodegenerative diseases (Orimaye et al., 2017), exposing widespread gender and ethnic biases in societies (Caliskan et al., 2017), and predicting large-scale trends in collective consumer behavior (Kallus, 2014). Because of this NLP has become a staple technology for everyday frequent tasks in most contemporary societies of the world. For instance, an English speaker with a smartphone can now easily get accurate information on many topics through a quick query to a virtual assistant, they can consult an online translation service to translate a foreign language web page with a click, and they can interact with many different machines and computers through simple speech commands.

These technological capabilities can be attributed to several developments over the last few decades: 1. the advent of deep learning methods, which allow for more effective creation of NLP systems from existing data (Goldberg, 2017), 2. the existence of standardized benchmark datasets and evaluation metrics (Wang et al., 2018; Hu et al., 2020), 3. the prestige afforded by the research community to researchers who improve upon these benchmarks, 4. the resulting large number of resources, be they computation, data, or ingenuity, that are poured into optimizing performance thereon. As both a theoretical and technical endeavor, NLP is experiencing an explosive increase: the annual conference of the Association of Computational Linguistics (ACL) received in 2000 less than 300 papers, growing in 2010 to slightly less than 1,000, to over more than 3,500 submissions in its 2020 edition. Largely as a result of this expansion of research effort, state-of-the-art systems have also achieved evaluation benchmark scores on par with human performance on a variety of NLP tasks such as question answering on English (He et al., 2021), or on automatic translation of news from German, Russian, and Chinese to English (Barrault et al., 2020).[2]

These upward slanting curves on standard benchmarks fail to show how uneven this development has been for all potential NLP users. Extensive research across NLP tasks have found systematic

[1] Data and code to reproduce the findings discussed in this paper are available on GitHub (https://github.com/neubig/globalutility).

[2] Although the significance of these parity claims has been disputed (Läubli et al., 2018).

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5486 - 5505

performance drops according to dimensions such as gender, racial identity, and language varieties, among others. The reasons for these biases are multifactorial and can be traced to virtually all stages in the process of NLP development, from the data used to train systems (Caliskan et al., 2017; Sap et al., 2019; De-Arteaga et al., 2019; Tatman, 2017; Tatman and Kasten, 2017; Buolamwini and Gebru, 2018; Raji and Buolamwini, 2019) to the very algorithms involved (Speicher et al., 2018; Bellamy et al., 2018; Adebayo et al., 2016). The growing awareness of these biases in NLP technologies brought by these studies, along with the development of novel metrics and tests to evaluate these disparities, have resulted in progressively more efficient and principled strategies to understand and mitigate them.

However, similarly systematic approaches are still lacking in one fundamental dimension of variation across individuals: their languages. Out of the over 6,500 languages spoken or signed in the world today (Hammarström, 2015), only a handful are systematically represented in academia and industry (Joshi et al., 2020; Yin et al., 2021). In spite of the aforementioned near-human results on translation or understanding of languages from the world's economic and political superpowers, the experience of any NLP practitioner is that, for the vast majority of languages, they fall far below such standards. Critically, the languages of the world showcase substantial variation in most domains of description, and in fact, the performance of language technologies has been shown to be sensitive to diverse aspects of the language under study, including morphology, word order, or phonological repertoire, as well as more mundane aspects like writing script or data availability (Arivazhagan et al., 2019; Tsarfaty et al., 2020; Xia et al., 2020; Muller et al., 2021). Hence, the transfer of NLP developments from one language to another is far from trivial, as it often means that building highly functional language technologies for any particular language is a non-automatic, costly, and technically challenging task.

Taking all these considerations together, and given that even the consequences brought by unequal NLP technologies across (racial, gender, socioeconomic) groups within the same nominal language are already substantial, there is a pressing need for measuring and understanding NLP performance inequalities across the world's languages. Here we develop novel estimates on how the utility afforded by NLP systems is distributed across individuals, languages, and tasks at an unprecedented global scale. These estimates allow us to identify which languages are systematically underserved by language technologies and could benefit the most individuals from focused technology development. We finally trace these inequalities to the societal, economic, and academic correlates of NLP systems' performance, shedding light on its latent causes, and indicate how our results favor specific evidence-based policies in research and development.

2 Methodology

2.1 Quantifying utility and demand

Our fundamental goal is evaluating the distribution of diverse representative language technologies (and their qualities) across the world's languages and their populations. Minimally, we would attempt to account for the patterns of association between the *demand* of language technologies and the *utility* they confer to users across languages. Thus, the first component of our analysis pertains quantifying the *utility* users in a given language l receive from a language technology. Ideally, such a measure would capture to what extent a given NLP system solves the specific problems an individual can pose to them - for instance, how successfully the user can obtain information from an automatically translated web page, or how satisfied the user is by a speech-based virtual assistant's execution of a series of verbal commands.

Intuitively, utility is associated with the nominal performance of the technology - a more performant system will allow the user to obtain a greater degree of utility. How "performance" is measured depends on the task (see Section 1). Since our purpose is to allow for comparisons, we define the utility of a task and language, u_l, as the corresponding performance normalized by the best possible performance afforded by such task, i.e.

$$u_l = \frac{\text{performance}_l}{\text{theoretical max performance}}$$

In cases where the best possible performance is undefined or technically unattainable, we take the empirical maximum as an estimate of the theoretical one and normalize by the best-performing language across all languages L, i.e. we replace the denominator in the above definition by $\max_{l' \in L}(\text{performance}_{l'})$.

Task	Description	Metric
Syntactic Analysis (DEP)	Infer syntactic dependencies between words in text	Labeled Attachment Score
Morphological Inflection (ING)	Produce an inflection given a lemma and morphological tags	Accuracy
Machine Translation (MT)	Translate text from a language into another	BLEU score
Speech Synthesis (TTS)	Produce speech on the basis of textual input	1-mel-cepstral distortion
Natural Language Inference (NLI)	Recognize entailment or contradiction between two sentences	Accuracy
Question Answering (QA)	Produce an answer for a textual query	Fscore

Table 1: NLP tasks evaluated in the present study, along with their corresponding performance metric.

Defining utility in this manner allow us to explore and contrast language technologies at the broadest scale, which is possible thanks to some necessary simplifying assumptions. As we pointed out before, not all users of the same language technology might benefit in the same manner given a fixed performance, and the relation between nominal performance and "true" utility might be complex and non-linear.[3]

With these caveats in mind, we further quantify the second component of our analysis, the *demand* for a language technology in each language l, d_l. We characterize d_l by taking into consideration demographic and linguistic perspectives. Under the first perspective, the demand for a given technology in a language is estimated to be proportional to the number of speakers of the language itself n_l ($d_l \propto n_l$). Under the second perspective, the demand across the approximately 6,500 languages of the world is identical ($d_l \propto 1$). These two alternatives as well as any intermediate combination of them can be simply parameterized through a single exponent τ,

$$d_l^{(\tau)} = \frac{n_l^\tau}{\sum_{l' \in L} n_{l'}^\tau}$$

where $\tau = 1$ correspond to a demographic notion of demand, $\tau = 0$ to a linguistic one, and $0 < \tau < 1$ is in between.

Equipped with these notions, we construct a simple family of global metrics (M_τ) revealing to what degree the global demand for language technologies is actually met:

$$M_\tau = \sum_{l \in \mathcal{L}} d_l^{(\tau)} \cdot u_l$$

[3]To give just one example, in machine-assisted human translation, the relationship between MT accuracy and productivity gain (directly associated with utility) is complex (Sanchez-Torron and Koehn, 2016).

M_τ has a number of intuitive properties we would like such a metric to have. M_τ is bounded between 0 and 1; 0 corresponds to a case where no-one benefits from a given language technology, whereas 1 would correspond to a situation where all languages enjoy perfect technology. Increasing the utility of a given language leads to an increase in M_τ, and the magnitude of this increase is influenced by both the size of the improvement and the demand in that language.

2.2 NLP tasks

We apply our measures of utility and demand to a set of diverse and representative NLP tasks, which are described below and summarized in Table 1.

The first three are tasks that technology users interact with directly in their everyday life, so that their output is already in a shape and form that is usable for most individuals. *Question Answering* (QA) consists of crafting a relevant answer to a question formulated in natural language, such as e.g. "what is the capital city of the Philippines?" or "why do dogs like bones?". This task is ubiquitous in online search or virtual assistants. *Machine Translation* (MT) is the task of translating from one language to another (e.g. from Tagalog to Estonian or from Japanese to Basque), and is typically used to facilitate inter-personal communication, information gathering, and e-commerce. *Text-to-speech* (TTS) is the task of rendering speech from textual input, which is used widely in spoken virtual assistants, car navigation systems, and is becoming a gateway for internet-of-things devices.

Next, *Natural Language Inference* (NLI) is a central task in AI and involves the evaluation of information presented in propositional format. More specifically, given a sentence called the "premise" (e.g. "the dog chewed a big bone"), NLI systems decide whether a separate sentence called the "hypothesis" is entailed by the premise (e.g. "the dog

gnawed at a bone"), negated by it (e.g. "the dog was sleeping"), or neither (e.g. "the dog likes bones"). While not a user-facing task *per se*, it measures the ability of NLP systems to adequately represent (and "understand") user queries.

Beyond these three (plus one) user-facing tasks, we also consider two more foundational linguistically-focused tasks, which often inform part of the pipelines of the user-facing tasks but which are rarely if ever encountered "in the wild" by language technology users. *Morphological Inflection* (Inflection) is the task of generating an inflected wordform given a lemma and a morphological specification, e.g. producing the third person singular form for "run": `run+3;SG`→`runs`. *Syntactic Parsing* under the dependency formalism (DEP) is the task of producing a syntactic parse of an input sentence, e.g. given the sentence "dogs like bones" specifying the "dogs" and "bones" are the subject and object of "like" respectively.

2.3 Correlates of NLP utility

Beyond the performance of individual tasks, we take a bird's-eye-view of the field of language technologies in general, as we analyze some of the correlates of the scientific production in NLP. In particular, we follow two broad guiding questions: (1) does the system of academic incentives promote the development of a more linguistically diverse NLP? and (2) is economic centrality or sheer demographic demand the best predictor of NLP technologies in any given language?

While a full understanding of the complex causal mechanisms binding society and NLP in general is outside of the scope of the present article, we set out to provide a first large-scale exploration of these matters by considering scientific publications appearing in major international NLP conferences as the basic units of science production. This simplification is not without challenges: for instance, some widely used language technologies are developed outside of the traditional scientific circuit based on proprietary technology, or they are published in local conferences, possibly in languages other than English.[4] In spite of this, studying scientific publications (and their correlates) allows us to evaluate transparent questions on the basis of publicly available data at a scale that is unfeasible for in-depth analyses.

[4]e.g. the Japanese NLP society's 2020 conference published 396 papers: https://www.anlp.jp/proceedings/annual_meeting/2020/

Therefore, we study the first question by determining whether the cumulative number of citations a paper receives is correlated with the number of languages it is associated with. We investigate our second question by finding the best predictive model of the number of NLP papers in any given language by contrasting two predictors: estimated number of users worldwide and approximate GDP associated with its users. We model these regression problems in a Bayesian generalized mixed effects framework (see Appendix B).

2.4 Data

We manually aggregate information on task performance and demand for the tasks summarized in Table 1 from a number of sources (we relegate many details to Appendix A, and give a high-level overview here). The data is taken from a combination of multilingual benchmarks, shared tasks and published results in NLP conferences including:

Question answering: We use data from the TyDi-QA (Clark et al., 2020), MLQA (Lewis et al., 2020), and SD-QA (Faisal et al., 2021) benchmarks and measure raw accuracy to calculate utility.

Machine translation: We aggregate scores from the WMT and IWSLT evaluation campaigns, and 50 studies from the last three years' ACL, EMNLP, and NAACL conferences, using BLEU (Papineni et al., 2002) as an accuracy metric.

Text-to-speech: We use data from the CMU Wilderness Project (Black, 2019) and use normalized negative mel-ceptral distortion (Kubichek, 1993) as an accuracy metric.

Natural language inference: We use results from the XNLI leaderboard (Conneau et al., 2018) and raw accuracy as the evaluation metric.

Syntactic parsing: We use the accuracies provided by UDPipe (Straka, 2018) and UDify (Kondratyuk and Straka, 2019) on the universal dependencies corpus (Zeman et al., 2017), with labeled attachment score as an accuracy metric.

Morphological inflection: We use results from SIGMORPHON workshops inflection shared tasks (e.g. (Vylomova et al., 2020)) measuring utility with exact-match accuracy.

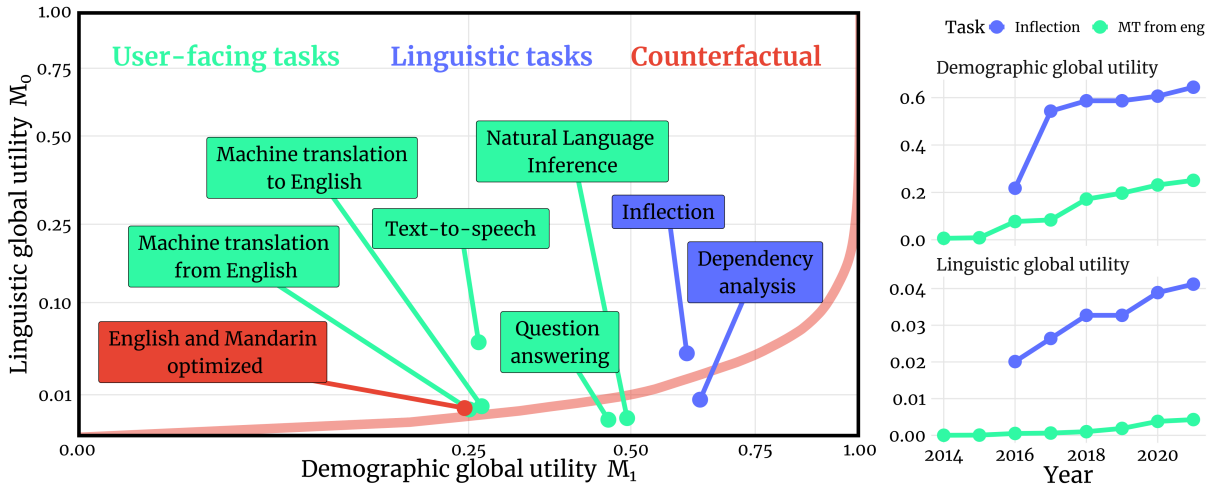


Figure 1: Left panel: linguistic and demographic global utility metrics for a number of language technology tasks. The red curve corresponds to the sequence where first the language with the largest number of users is set to utility 1, then the second, and so on. Right panel: recent historical progression of two language technology tasks: Inflection and Machine Translation from English.

Demographic and linguistic information necessary for the estimation of demands were obtained from a variety of sources, including Ethnologue, Glottolog, and the World Trade Organisation. For most tasks, the number of first-language speakers is used to measure demand, but for MT we estimate the need for translation between two languages based on economic indicators of interaction between countries, and the language-speaking populations within the countries where the language is spoken.

3 Results and Analysis

3.1 General observations

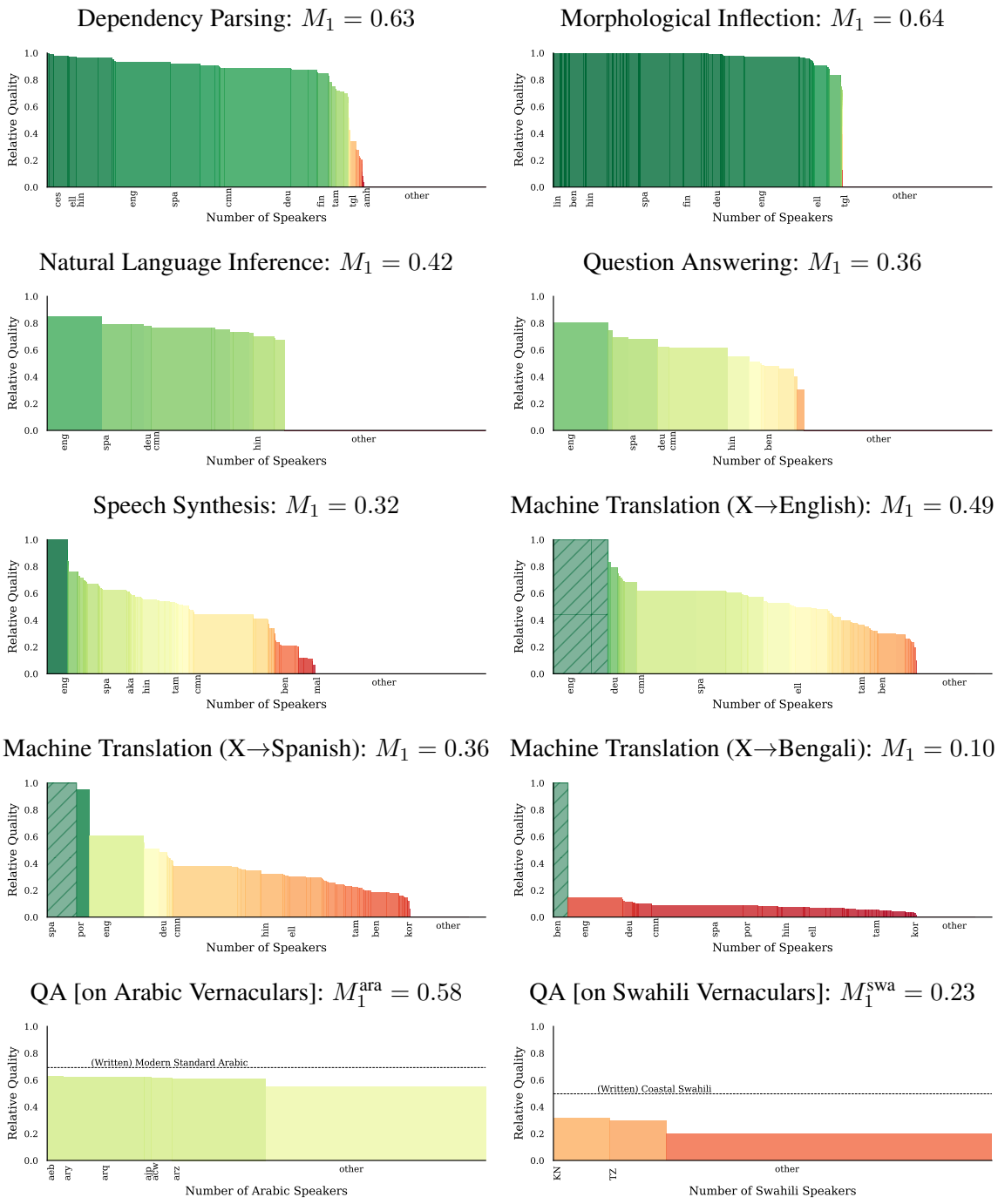
Figure 1 presents an overview of our main findings. Unsurprisingly, most NLP tasks we focus on fare substantially better when utility is measured demographically rather than linguistically.

Text-to-speech synthesis is the task with the most linguistic coverage: the published results (due to a single study (Black, 2019)) cover more than 630 languages (or about 10% of the world's languages). However, for the vast majority of these languages the measured quality of the generated speech is about half as good as the exceptionally good English system (Ren et al., 2021). The next most linguistically diverse tasks are those regarding morphosyntactic analysis, i.e. morphological inflection and dependency parsing, which have been evaluated over 140 and 90 languages respectively. For these more esoteric tasks which do not necessarily convey direct utility to a downstream user, the majority of the systems are in general very good.

Natural language inference (NLI; a representative natural language understanding task) and question answering (QA) lie on the opposite side of the spectrum: the established benchmarks have only focused on up to 15 and 17 languages respectively, leading to very low scores on the linguistic axis.

In Figure 1 (right panel) we observe the progress of the utility metrics in tasks for which we had access to comparable data across a span of the last 7 years. The extensive efforts of the UniMorph project (Kirov et al., 2018) to cover as many languages as possible are visible in the "Inflection" plot, with significant improvements over time. On the other hand, the machine translation field is still in the process of ramping up following demographics and/or socioeconomic priorities, with improved linguistic coverage over the years.

The granularity of these findings can be increased on the basis of available data. Figure 2 additionally presents demographic utility across language populations for all tasks. The visualization allows for identification of ostensive gaps in received utility. The two bottom plots of Figure 2 display our metrics over speakers of a single language, based on question answering results for different spoken Arabic and Swahili lectal varieties (Faisal et al., 2021). This analysis shows that utility differences are small between Arabic vernaculars although these systems still lag behind the systems for Modern Standard Arabic, while the utility level of Coastal Swahili speakers in Tanzania is about 10% lower than that for speakers in Kenya.



acw: Hijazi Arabic, aeb: Tunisian Arabic, ajp: South Levantine Arabic, aka: Aka, amh: Amharic, arq: Algerian Arabic, ary: Moroccan Arabic, arz: Egyptian Arabic, ben: Bengali, ces: Czech, cmn: Mandarin Chinese, deu: High German, ell: Greek, eng: English, fin: Finnish, hin: Hindi, kor: Korean, lin: Lingala, mal: Malayalam, por: Portuguese, spa: Spanish, swa: Swahili, tam: Tamil, tgl: Tagalog.

Figure 2: Illustration of our metric on demographic-focused utility ($\tau = 1$) on various NLP tasks.

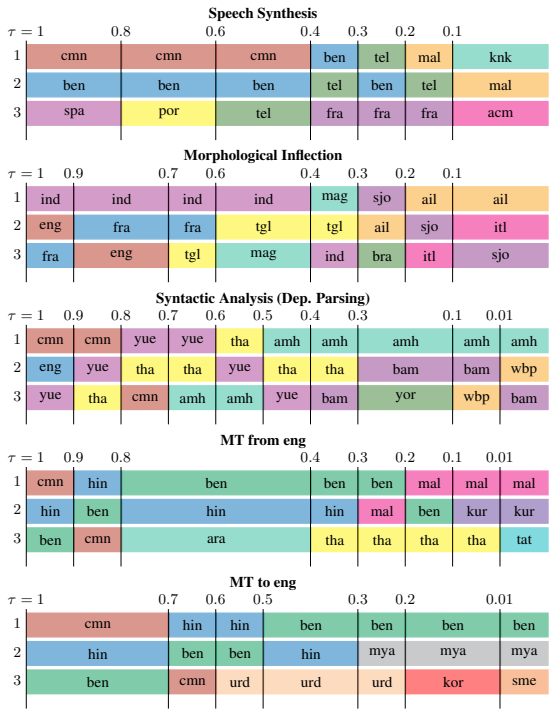


Figure 3: Priority languages (top-3 shown) change with different balancing of demographic and linguistic utility, with focus shifting from populous languages e.g. Mandarin (cmn) and Hindi (hin) to more under-served languages e.g. Kuranko (knk), Bambara (bam), Tatar (tat), or Aimele (ail).

3.2 Priorities in NLP development

Given the current snapshot of NLP systems, we could ask which languages will lead to the largest global utility improvement. The relative importance of linguistic vs. demographic demands determines the priority ranking, as it can be observed in Figure 3 for a sample of five tasks. Improving on the demographic-focused utility entails a greater emphasis on Mandarin Chinese, Hindi, Spanish, and other populous languages that are generally well-served by current technologies. Balancing linguistic and demographic considerations leads to prioritizing a more diverse set of languages, mostly Asian and African languages like Amharic, Bambara, Bengali, Thai, or Yoruba, which are both populous and under-served, along with also large but severely under-served languages like Kurdish, Urdu, and Oromo. Further emphasis on linguistic utility would lead to prioritization of indigenous and potentially endangered languages of small communities like Aimele, Itelmen, North Sami, or Warlpiri, which are currently largely ignored by NLP research (Bird, 2020).

3.3 The role of society, economy, and academia

Now we turn to our large-scale analysis of NLP publications. First, this reveals that a substantial proportion of publications do not even describe in a clear and unequivocal manner the language (or languages) they are dealing with (Bender, 2011). Given the current prevalence of English of a language of study in NLP, in most cases, the lack of an explicit reference to a particular language entails the system deals with English exclusively.

This reflects a more deep-seated issue reflected in the citation of papers over time. Independently of publication venue, year, or subfield of NLP research, the number of languages a publication deals with is not predictive of how many citations it will accrue over time (see Figure 4, top right panel). In other words, if citations can be regarded as a proxy for academic incentives, scientists and developers are presented with little to no additional academic reward when tackling data, problems, or tasks involving more than one language.

This naturally leads to the question of what explains the production of language technologies across languages to start with, which will necessarily involve agents, mechanisms, and data, outside of the scope of NLP publications themselves. Nevertheless, in order to contribute to this investigation, we determined whether approximate measures of economic centrality or number of language users were better predictors of sheer number of papers published for any given language (see Appendix C). While both variables are substantially collinear, we find that approximate GDP (rather than number of users) leads to a substantially smaller prediction error of number of published papers.

4 Discussion

Our study, covering diverse NLP tasks and types of evidence, makes apparent the immense inequality in the development of language technologies across the world's languages. After English, a handful of Western European languages dominate the field -in particular German, French, and and Spanish- as well as even fewer non-Indo-European languages, primarily Chinese, Japanese, and Arabic. Our preliminary investigation suggests it is the economic prowess of the users of a language (rather than the sheer demographic demand) what drives the development of language technologies.

In spite of this, for some tasks (such as In-

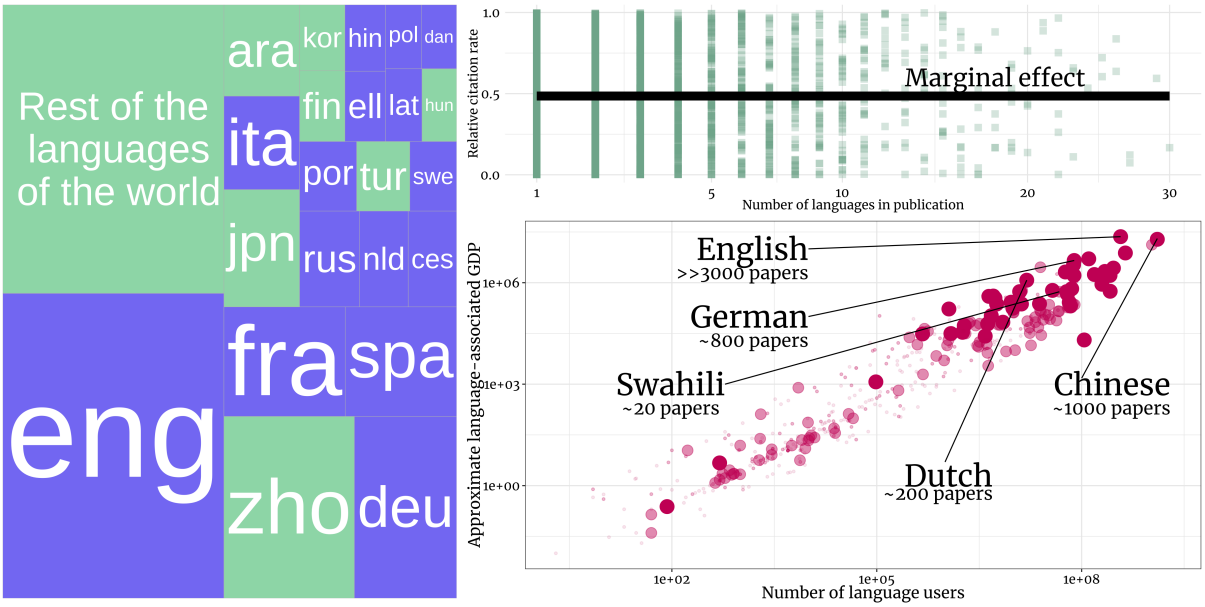


Figure 4: Left panel: treemap of the number of NLP publications per language (with area proportional to the number). Right top panel: Relative citation rate vs number of languages in the publication. Right bottom panel: Number of publications according to number of language users and approximate GDP. Point size and transparency scales with number of publications. eng: English, zho: Chinese, deu: German, fra: French, spa: Spanish, jpn: Japanese, rus: Russian, nld: Dutch, ces: Czech, por: Portuguese, tur: Turkish, swe: Swedish, ita: Italian, fin: Finnish, ell: Greek, lat: Latin, hun: Hungarian, ara: Arabic, kor: Korean, hin: Hindi, pol: Polish, dan: Danish.

flection) there is an encouraging trend of both demographic- and linguistic-utility improving year-over-year. This is due to the nature of the task; reasonably accurate solutions can be achieved through small but highly-curated data. Since linguistic expertise on the languages of the world is, naturally, globally distributed, the main hurdle these tasks face is to pool such expertise under the premise of a common technical goal. In this respect, relatively low-cost and bottom-up actions that gather experts to work on specific NLP tasks (such as Universal Dependencies and UniMorph) have succeeded in accelerating the cross-linguistic development of language technologies. These prosper mainly on the basis of academic incentives, as those individuals or groups who contribute data and/or expertise are rewarded with individual publications or co-authorship in collective publications. Many of these contributions - which do not necessarily involve hefty resource investments but instead linguistic expertise - are markedly different from the typical publications in language technologies.

However, these more esoteric tasks are tenuously associated with those that users are more likely to interact with, such as Machine Translation or Speech Synthesis. User-facing tasks all have in common a tight dependency on computational resources and large data, which in turn hinge on substantial financial means. In a context of pressing user needs across multiple populations and languages, we submit that future developments on policies aimed at furthering cross-linguistic technologies would benefit from clear (and possibly standardized) metrics that assist in streamlining complex decisions regarding resource allocation. Our measures of global coverage fulfill that role, and help identifying large but currently under-served languages. While we do not attempt to supplement the necessary in-depth evaluation of the need of each individual group and language, they provide a common ground for coordinating global efforts across heterogeneous actors.

In addition, we would like to reiterate that our work here has necessarily made a large number of simplifying assumptions to even attempt to quantify disparities in language technology utility on a global scale. These most notably involve simplifying assumptions regarding the measurement of demand (based on native-speaker population and/or economic indicators) and the measurement of utility (based on simple accuracy metrics). Future work may further clarify these assumptions, making more accurate estimates of true user demand on a technology-by-technology level, or more accurately clarifying the relationship between standard accuracy metrics and the utility derived by users.

Acknowledgements

This work was supported by NSF Award 2040926.

References

Julius A Adebayo et al. 2016. *FairML: ToolBox for diagnosing bias in predictive modeling*. Ph.D. thesis, Massachusetts Institute of Technology.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana. Association for Computational Linguistics.

Gopala Krishna Anumanchipalli, Kishore Prahallad, and Alan W Black. 2011. Festvox: Tools for creation and analyses of large speech corpora. In *Workshop on Very Large Scale Phonetics Research, UPenn, Philadelphia*, page 70.

Mihael Arcan, Maja Popovic, Paul Buitelaar, et al. 2016. Asistent–a machine translation system for slovene, serbian and croatian. In *Proceedings of the 10th Conference on Language Technologies and Digital Humanities, Ljubljana, Slovenia*.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv:1907.05019.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Regina Barzilay and Min-Yen Kan, editors. 2017. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv:1810.01943.

Emily M Bender. 2011. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975. IEEE.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91.

Paul-Christian Bürkner. 2017. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1):1–28.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. Stan: a probabilistic programming language. *Grantee Submission*, 76(1):1–32.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in ty pologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the*

CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2018. Ethnologue: Languages of the world. twenty-second edition. SIL International.

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoav Goldberg. 2017. Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10(1):1–309.

Iryna Gurevych and Yusuke Miyao, editors. 2018. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia.

Harald Hammarström. 2015. “Ethnologue" 16/17/18th editions: A comprehensive review. *Language*, 91(3):723–737.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *Proceedings of the International Conference on Learning Representations*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors. 2019. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China.

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.

Nathan Kallus. 2014. Predicting crowd behavior with big public data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 625–630.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kevin Knight, Ani Nenkova, and Owen Rambow, editors. 2016. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795.

Anna Korhonen, David Traum, and Lluís Màrquez, editors. 2019. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy.

R Kubichek. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128. IEEE.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a

case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796.

M Paul Lewis, Gary F Simons, Charles D Fennig, et al. 2009. *Ethnologue: Languages of the world*, volume 16. SIL International.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Sylvester O Orimaye, Jojo SM Wong, Karen J Golden, Chee P Wong, and Ireneous N Soyiri. 2017. Predicting probable alzheimer's disease using linguistic deficits and biomarkers. *BMC bioinformatics*, 18(1):1–13.

Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors. 2017. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. ACL*, pages 311–318.

Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435.

Yi Ren, Chenxu Hu, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2021. Fastspeech 2: Fast and high-quality end-to-end text-to-speech. In *Proceedings of International Conference of Learning Representations (ICLR)*.

Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors. 2018. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium.

Marina Sanchez-Torron and Philipp Koehn. 2016. Machine translation quality and post-editor productivity. In *Proceedings of AMTA*, volume 2016, page 16.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. 2018. A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2239–2248.

Milan Straka. 2018. Udpipe 2.0 prototype at conll 2018 ud shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Rachael Tatman and Conner Kasten. 2017. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, pages 934–938.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27(5):1413–1432.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Marilyn Walker, Heng Ji, and Amanda Stent, editors. 2018. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors. 2020. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online.

Mengzhou Xia, Antonios Anastasopoulos, Ruochen Xu, Yiming Yang, and Graham Neubig. 2020. Predicting performance for natural language processing tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8625–8646, Online. Association for Computational Linguistics.

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

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Materials

Publication data We rely on papers available through the Anthology of the Association of Computational Linguistics[5] which hosts more than 60 thousand papers from all major NLP conferences. We rely on Semantic Scholar (Ammar et al., 2018) for citation information.

We make the working assumption that a mention of a language in a research paper likely entails that the underlying research involves this language. We follow an automatic pipeline for finding language mentions in a paper, which starts by converting the paper PDF to a machine-readable format. We then search within the paper for any mention of a language's English name(s), its endonym, as well as its ISO or Glottolog code. We then apply a post-processing step to ensure the precision of this pipeline as our simple text-based search is prone to false positives for languages whose names match common English words (e.g. She, Male, Label, Even, The, Are), common placenames (e.g. Colorado, Nara, Sydney), parts of author names (e.g. Su, Kim, Dan, Ali, Rama), or mathematical notation (e.g. Dji, Dii).

In addition, we enrich each publication by imputing its research area. There were 16 research areas identified, based on the ones represented at recent major NLP conferences (specifically starting with the 2019 version of EMNLP, and removing some of the areas that were unique to that conference). For each area, we identified 1-6 publication venues from the ACL Anthology, where more venues were

[5] https://www.aclweb.org/anthology/

chosen when each venue had relatively few publications. Based on the abstracts of papers from each of these venues, we trained a bag-of-words classifier using the linear support vector machine implementation in scikit-learn[6], and applied this classifier to the abstracts of the papers we wanted to classify. Necessary data and code to reproduce these results are released in the supplementary material.

Data Sources and Metrics for Utility The majority of NLP research relies on automatic evaluation metrics over datasets annotated with gold-standard outputs. The advantage of this approach is that it allows consistent comparisons between systems and a seamless evaluation of progress on a specific evaluation set. On the other hand, there is no guarantee that even statistically significant improvement on an automatic metric translates to improvements on user-perceived utility. Nevertheless, the reality is that virtually all published NLP research reports automatic evaluation metrics, with only a tiny fraction diverging from the norm by e.g. using human evaluations.

Our analysis assumes that all named languages have standard versions that are comprehensible and acceptable to all members of the population identified as "speakers" in our sources. However, we have the demographic information necessary for more fine-grained analysis in only a handful of languages. While this assumption is certainly an oversimplification, we nevertheless believe it does not detract from our paper's arguments.

For a completely fair comparison across languages, one would ideally compute automatic metrics over the same or an equally representative evaluation set. For our language understanding case study this requirement is satisfied, as the XNLI 15 language test sets are translations of the same evaluation set. Utility in this case, where the evaluation metric m is accuracy, will be equal to the accuracy for each language's l test set: $\text{utility}(l, m) = m_l$.

Natural language understanding results are sourced from the XNLI leaderboard (Conneau et al., 2018), which contains test datasets with premise-hypothesis pairs in 15 languages.

For question answering (QA) we aggregate results from two established multilingual benchmarks, namely TyDi-QA (Clark et al., 2020) and MLQA (Lewis et al., 2009). Both benchmarks focus on extractive question answering, i.e. finding the text span of a given document that answers, if possible, a given question. We also include SD-QA (Faisal et al., 2021) for additional dialectal breakdown for some of the TyDi-QA languages. The benchmarks jointly cover 17 languages. We keep the highest results for languages that are shared between the two datasets (English and Arabic). For this task we equate utility with test set F-score, a measure that meaningfully combines precision and recall of the retrieved answer span.

For machine translation, we collected more than 500 published MT results from all WMT and IWSLT evaluations, as well as more than 50 MT studies from the last three years' ACL, EMNLP, and NAACL conferences (Barzilay and Kan, 2017; Gurevych and Miyao, 2018; Palmer et al., 2017; Riloff et al., 2018; Knight et al., 2016; Walker et al., 2018; Korhonen et al., 2019; Inui et al., 2019; Webber et al., 2020). In the machine translation field the most popular evaluation metric is BLEU (Papineni et al., 2002). In our MT case studies we estimate utility based on a normalized version of BLEU, such that for translation from s to t with $\text{BLEU}(s, t)$ over an established test set, we have $\text{utility}(s, t, \text{BLEU}) \approx \frac{\text{BLEU}(s,t)}{Z}$. The normalizing factor $Z = \max_{\mathcal{L}\times\mathcal{L}} \text{BLEU}$ is equivalent to the largest reported BLEU, which we equate to the largest attainable utility at the snapshot of interest. In all our MT case studies we use $Z = 70$, which is the BLEU score reported for translation between Serbian and Croatian (Arcan et al., 2016).

For text-to-speech synthesis, we relied on results from the CMU Wilderness project (Black, 2019), which builds TTS voices with FestVox (Anumanchipalli et al., 2011), and compared them to the English system of (Ren et al., 2021). The quality of the synthesized audio is evaluated using mel-cepstral distortion (Kubichek, 1993, MCD) a distortion measure that compares synthesized examples with originals (lower is better). Each MCD of x_l for a language l was converted to a relative utility score by applying the transformation $\frac{x_{\max} - x_l}{x_{\max} - x_{\min}}$, where $x_{\max}$ and $x_{\min}$ correspond to the highest (worst) and lowest (best) observed MCD scores across all languages.

For syntactic analysis through dependency parsing, we relied on results from two state-of-the-art systems, UDPipe (Straka, 2018) and UDify (Kondratyuk and Straka, 2019). The systems are typically evaluated using two measures, Unlabeled and Labeled Attachment Score (UAS and LAS), which measure the overlap between human-created and

[6] https://scikit-learn.org/stable/

automatically-produced syntactic trees, excluding punctuation. For our metrics we use LAS, which considers the semantic relation (e.g. Subj) used to label the attachment between two words.

The results on morphological inflection were taken from the findings of the corresponding shared tasks that have been taking place as part of the SIGMORPHON workshop for the past 5 years (Cotterell et al., 2016, 2017, 2018; McCarthy et al., 2019; Vylomova et al., 2020). The systems are evaluated using exact-match accuracy over a predefined test set in each language, simply comparing the correct inflected form with the system's output.

Population Demand We compile population statistics from various sources. We rely on Ethnologue (Eberhard et al., 2018) for language population statistics. We take special care when computing population statistics over macro-languages (e.g. Arabic, Chinese) and languages commonly spoken by L2 speakers (e.g. English) or across multiple dialects (e.g. for Spanish or Portuguese), aggregating populations across all variants.

Economic Indicators for Demand We aggregate economic information on international trade, as provided from the World Trade Organisation (WTO) through the World Integrated Trade Solution.[7] Since each language community can be geographically associated with a member nation of WTO, we can then estimate economic indicators for and between language communities.[8]

In a monolingual setting, we rely on the most recent GDP estimates, associated with each language community. For example, the 1.7 million Nahuatl speakers represent about 1.3% of Mexico's population, and thus the final GDP associated with the Nahuatl language will be 1.3% of Mexico's GDP.

Modeling demand in a bilingual setting (across two languages) is also feasible using economic indicators. For instance, the amount of trade between two language communities could be used to approximate the need for translation between the two. Specifically, if we use the normalized import volume per language community then we can estimate demand for an $s \rightarrow t$ translation system as $\text{demand}(s,t) \propto v^{\text{import}}_{s \rightarrow t}$ such that $\sum_{s \in \mathcal{L}} v^{\text{import}}_{s \rightarrow t} = 1$.

[7] https://wits.worldbank.org/

[8] Our conclusions and analyses based on WITS data are the responsibility of the authors and do not represent the opinion of the WTO.

Take the Azerbaijani language as an example: Azerbaijan's imports mainly come from the Russian Federation (16.8%), Turkey (14.7%), China (11.2%), the US (8.5%), Ukraine (5.5%), and Germany (5.5%).[9] Hence, we can assign a proportional weight to model demand for translation from Russian, Turkish, Chinese, English, Ukrainian, and German into Azerbaijani respectively. One could equivalently use the normalized volume of exports instead.

This is only straightforward to compute in cases where a language is easy to map to a specific country. In cases of languages that are commonly used across many countries e.g. German (which is the main language in both Germany and Austria) or macro-languages spoken in larger regions of the world, we combine the weights accordingly in order to jointly model the demand for the whole language community.

Table 3 presents the top-15 translation pairs based on demand estimated from economic indicators, namely the import (and export) partner share of the target (source) language. We note that this ranking does not take underlying populations into account, using only the *percentage* of demand for each language community. Several entries in Table 3 are language pairs that are rarely, if ever, studied in MT case studies, like Belarusian-Russian, Mongolian-Mandarin Chinese, Albanian-Italian, or Russian-Armenian.

B Methods

Predicting Utility on Unseen Languages/Pairs One of the main disadvantages of using solely published results for estimating quality and, hence, utility, is the lack of evaluations on all languages or language pairs. Furthermore, not all languages or pairs are consistently evaluated on newly developed models. To counter this issue, we propose a more comprehensive approach which attempts to predict the expected quality/utility over languages or language pairs unseen in the collected literature.

A naive approach is to make the approximation that utility on any unseen language is 0. However crude, this could be a valid assumption in many cases: consider the example of a language understanding system trained on all languages that appear in Wikipedia. Such a system, without proper

[9] Source: https://wits.worldbank.org/CountryProfile/en/Country/AZE/Year/2017/TradeFlow/Import

modifications, would not be able to handle input in Yupik or Dhivehi (Maldivian), since these languages are not represented in Wikipedia and they use different writing systems than any other language. Note that, in such a case, for a language understanding system evaluated over a classification task as in a language understanding setting, the expected utility is not 0, but is rather the expected quality of random outputs (33% in the case of three-way classification).

Future work could make use of models explicitly trained to predict the accuracy (or other metrics) of NLP models on unseen languages or language pairs, such as the ones proposed by (Lin et al., 2019) or (Xia et al., 2020).

Estimating MT quality with pivoting In the case of machine translation, pivoting is a viable approach for producing translations between any arbitrary language pair, as long as the intermediate systems exist. Even if no published results exist on translation from German to Chinese, it is unreasonable to assign an expected utility of 0 to such a MT system, since there exist high-quality German-English and English-Chinese systems.

In the case of cascaded systems, though, estimating utility requires a careful approach, due to error propagation. Consider a system A with accuracy 80% and a system B with accuracy also 80%. A cascaded system where the output of system A is provided as input to system B will have an expected accuracy 64%, not 80%.

An important point is that there is no reason for pivoting through a single language. Consider the example of Catalan to Chinese translation. A path from Catalan to Spanish, to English, to Chinese might have a yield a higher estimated utility from a single-language pivoting path, since its components are of higher quality.

We devise a method that allows us to generalize this notion in order to find the highest estimated utility for every language pair. We construct a weighted directed graph $\mathcal{G}$=(V, E) with each node $v \in V$ representing a language. The weighted directed edge $e_{s \to t}$ between nodes s and t will have a weight equal to the highest reported normalized BLEU score on translation from s to t. If no results have been published on this language pair, we set the weight of that edge to 0.

With graph $\mathcal{G}$ in hand, as long as a path from nodes s to t exists, we can estimate the expected normalized BLEU of $s - t$ translation as the maximum cumulative (multiplicative) weight over any path from s to t. If a path does not exist, then the estimation is 0. This is possible in cases where a language is reported as only source or only target in the literature; for example, Greek (ell) only appears as a source in a single study (reporting Greek–English translation results) which allows us to estimate Greek–X utility by pivoting through English, but we cannot produce estimates for X–Greek. Table 4 presents translation pairs were our estimated utility (normalized BLEU score) is higher than the published results.

C Bibliometric Analysis

Analysis of Citations To each publication we associate its citation percentile relative to its year and event. We analyze normalized citations (C) through Bayesian generalized additive mixed effects models implemented in R with brms and Stan (Bürkner, 2017; Carpenter et al., 2017) We utilize default weakly informative priors for all parameters and we run four MCMC chains for each model which in all cases achieved convergence. The distribution of C is described through a beta distribution, of which its expected value is given by

$$\mathbb{E}[C] = \text{logit}(f(L) + \alpha_A + \beta_A \cdot L) \quad (1)$$

where $f(L)$ is a smooth function (on the basis of thin plate splines) depending on the number of languages dealt with in the paper (L), and α_A and β_A are random intercepts and slopes according to each area, respectively. In order to evaluate the support in favor of $f(L)$, we compared the leave-one-out (LOO) performance of this model against a counterpart without this term,

$$\mathbb{E}[C] = \text{logit}(\alpha_A + \beta_A \cdot L) \quad (2)$$

The difference in expected log pointwise predictive density (which serves to inform model selection, (Vehtari et al., 2017)) between the two models is -0.9 (SE=0.6), which implies there is no major performance difference between the two.

Analysis of Number of Publications We determine the total estimated number of papers in which each language l was involved (P_l). The resulting distribution has a large concentration of zero values, so we opt to model this through a zero-inflated negative binomial distribution. We focus on two parameters: the expected value of the number of publications ($\mathbb{E}[P]$) and the mixture probability (π).

In both cases, we fit models considering three possibilities: (1) A smooth (thin plate spline) function of the log-GDP, (2) a smooth (thin plate spline) function of the log-number of speakers, and (3) a fixed parameter. This leads to evaluating 9 models through a LOO criterion. The model that involves (1) for both parameters displays the best overall performance (see SI).

D Machine Translation Case Studies

We use this section to expand on the discussion of MT case studies.

Translation involving English Since translation involves two languages and language communities, there are two natural ways for a speaker to receive utility from a MT system: either by being the *source* (with their language being translated into another) or by having another language translated into theirs (*target*). We disentangle the two by only using each one at a time for our utility calculations.

Utilities based on demographics for both settings are similar, with $M = 0.25$ (from English) and $M_1 = 0.27$ (to English). Since published results only cover 101 languages, the linguistic diversity scores are much lower, with M_0 around 0.005.

Translation among all languages We extend our study on translation among all languages (still maintaining the distinction between a language used as source or target). We base our estimates for utility on any reported results, as well as on accuracy estimates based on a pivoting approach. Briefly outlined, our pivoting estimation approach finds the best performing translation path for language pairs without reported results, i.e. since no studies report translation accuracy when translating from Greek to Chinese, we find that among all possible translation paths, translating from Greek to English and from English to Chinese yields the highest expected accuracy. We outline the process in the Materials and Methods section.

Perhaps unexpectedly, the best (and often only) pivot is English in almost all cases. As a result, the final utility for a language X is very much dependent on the utility of the X-Eng (or Eng-X) systems. This is reflected by our scores for averaged by demographics and languages being very similar to the ones when we only focused on English. Nevertheless, the differences between scores for different languages are stark: the demographic-averaged utility for populous, well-studied languages like German ($M_1 = 0.356$), Chinese ($M_1 = 0.232$), or French ($M_1 = 0.309$) is almost double than underserved ones like Bengali ($M_1 = 0.148$), isiXhosa ($M_1 = 0.156$), Amharic ($M_1 = 148$), or Burmese ($M_1 = 0.092$). Figure 5 visualizes the different scores for translation from 24 languages under the demographic focus ($\tau = 1$).

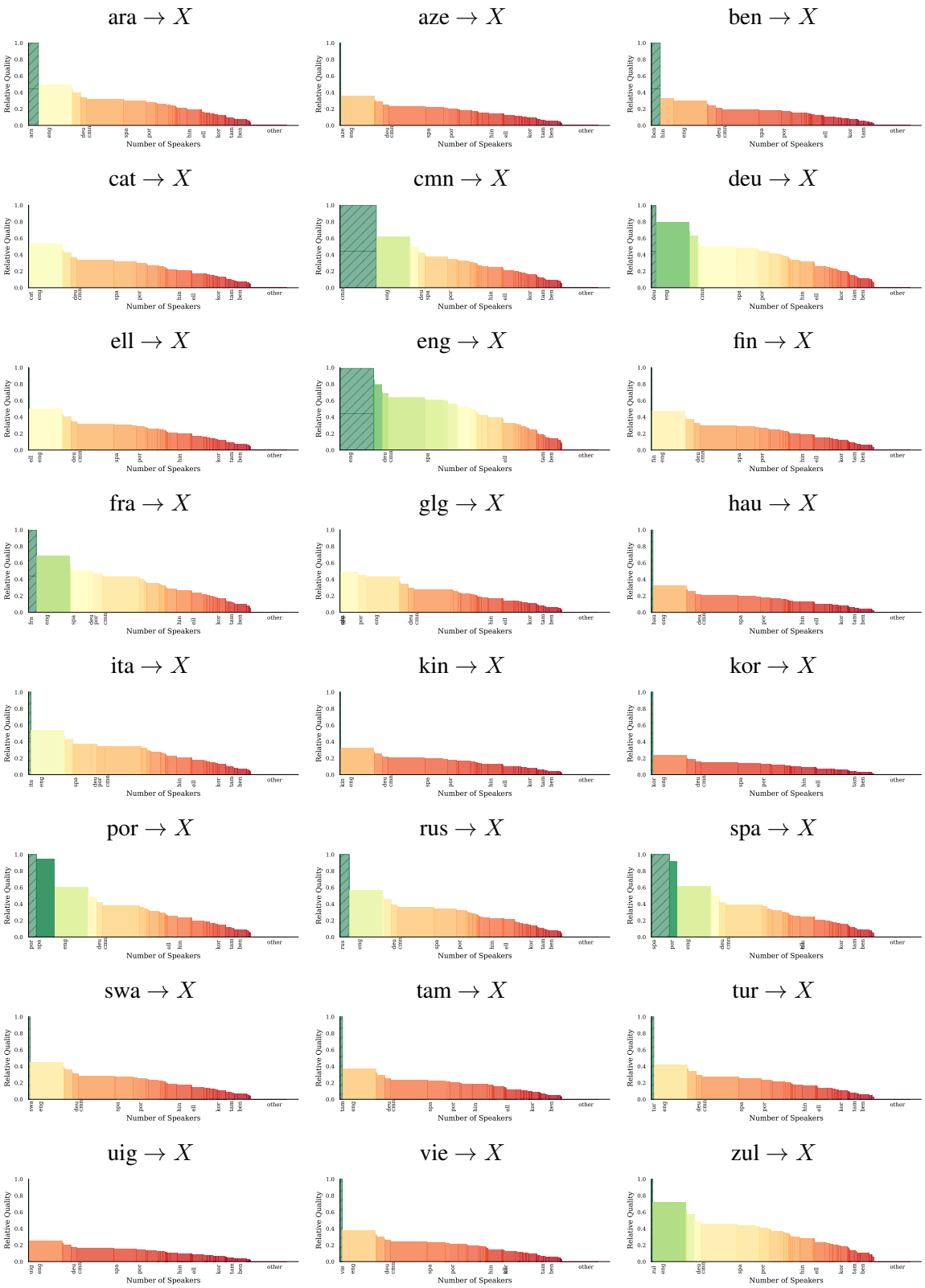


Figure 5: Visualization of our measure on translation from 24 diverse languages.

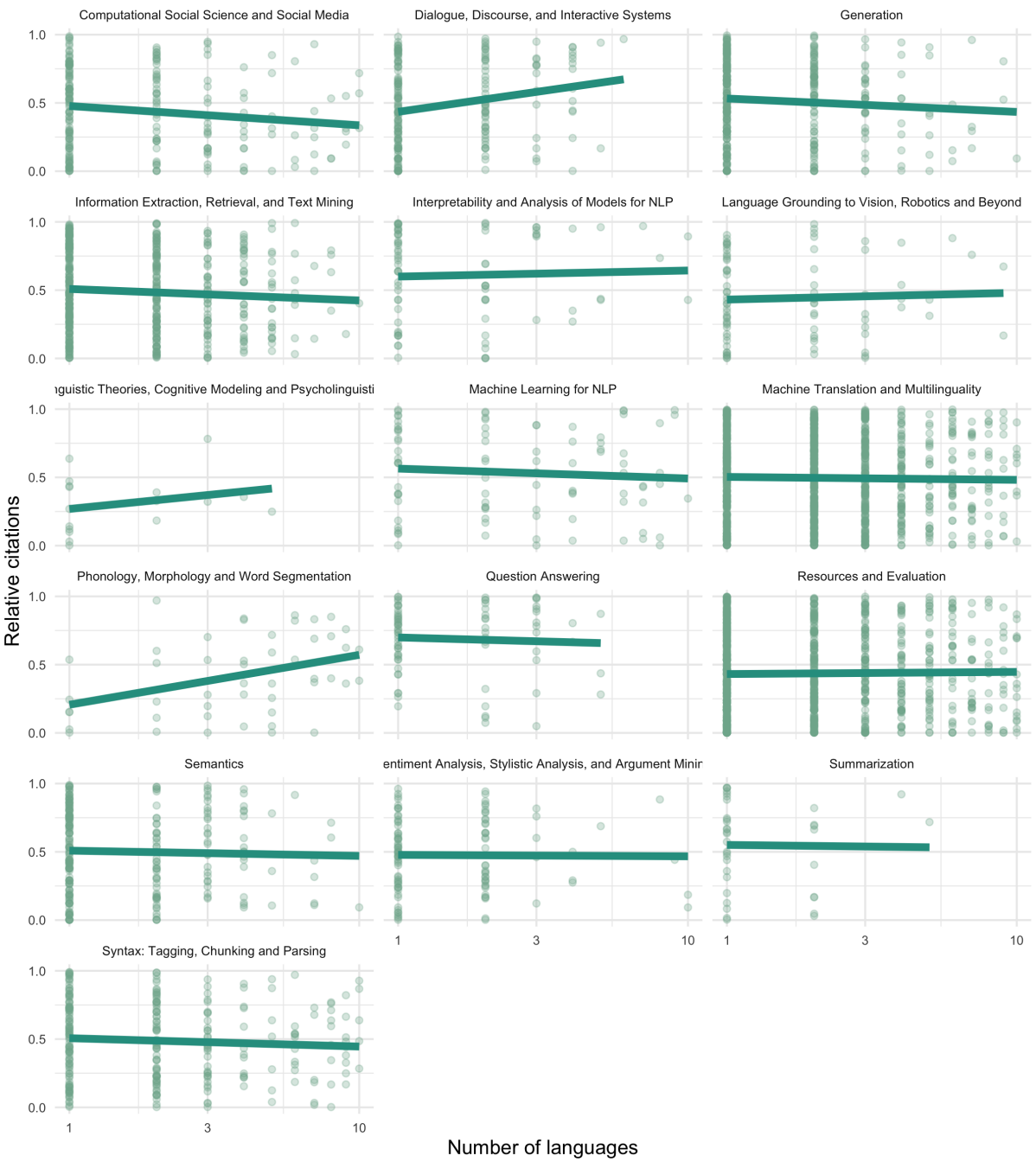


Figure 6: Cumulative citations vs number of languages in publications according to topic

rank	Lang.	$\mathrm{pop}_{-\mathrm{eng}}$ (M)	Number of Studies X–eng/eng–X
1	cmn	908.8	16/ 4
2	spa	358.8	5/6
3	hin	299.5	3/1
4	ben	232.8	2/0
5	por	207.7	3/3
6	ara	205.4	9/6
7	rus	145.6	9/6
8	jpn	128.0	7/4
9	swa	89.2	1/1
10	msa	80.3	2/0
11	kor	77.3	4/0
12	vie	76.0	4/6
13	mar	73.0	2/0
14	tam	72.0	2/0
15	tur	65.9	9/4
16	guj	48.3	1/1
17	fra	47.1	12/17
18	ind	43.4	2/0
19	ita	42.8	8/6
20	urd	35.0	2/0
21	mya	31.4	2/0
22	mal	30.7	0/0
23	deu	30.4	25/33
24	orm	28.0	1/0
25	uzb	27.9	0/0
26	ukr	27.3	3/1
27	pol	25.0	2/0
28	aze	19.5	5/2
29	sin	17.6	1/1
30	ron	16.8	13/11

Table 2: Machine Translation research interests on to and from English do not match our population-based demand model.

Rank	Based on	
	Imports	Exports
1	rus–bel	bel–rus
2	rus–kaz	mon–cmn
3	rus–hye	sqi–ita
4	rus–mon	hye–rus
5	rus–cmn	tgl–jpn
6	spa–som	nep–hin
7	hin–nep	aze–ita
8	ita–sqi	srp–bos
9	lit–lav	lav–lit
10	rus–aze	msa–jpn
11	cmn–mya	lit–rus
12	rus–fin	mya–cmn
13	rus–ukr	est–fin
14	cmn–tha	bos–hrv
15	jpn–tgl	kat–rus

Table 3: Top-15 translation pairs based on demand estimated from economic indicators (import (export) partner share of the target (source) language).

Language Pair	BLEU Score Estimated	BLEU Score Published	Pivot
slv–srp	37.09	25.45	eng–hrv
eng–nep	10.56	6.8	guj–hin
eng–hrv	60.80	42.15	srp
eng–hin	13.78	12.5	guj
hrv–eng	50.42	48.07	srp
ron–deu	29.36	18.4	eng
ron–fra	33.98	26.53	eng
ces–rus	17.56	16.2	eng
ces–deu	23.36	19.3	eng
ces–fra	27.04	18.1	eng
ita–deu	26.08	19.85	eng
rus–ces	18.19	14.4	eng
pol–ces	9.90	7.2	eng
nld–deu	25.0	21.06	eng
heb–fra	27.41	23.25	eng
srp–slv	52.09	35.39	hrv
deu–ron	27.25	16.27	eng
deu–ces	25.19	20.1	eng
deu–ita	28.42	18.56	eng
deu–nld	26.48	20.31	eng
deu–fra	44.27	37.3	eng
fra–ron	23.52	19.3	eng
fra–ces	21.73	13.7	eng
fra–heb	18.88	13.54	eng
spa–ces	17.83	15.2	por–eng
ara–fra	26.83	25.07	eng
slv–hrv	55.64	40.44	eng–srp

Table 4: Translation pairs with a pivoting estimated utility (BLEU score) higher than the published result.

Parameter Negbinomial	Parameter Zero-inflated	ELDP difference	SE
log-GDP	log-GDP	0	0
log-GDP	log-Users	-20.2	6.3
log-Users	log-GDP	-31.9	9.8
log-Users	log-Users	-69.8	13.2
log-GDP	Fixed	-87.9	15.1
log-Users	Fixed	-125.2	17.4
Fixed	log-GDP	-263.3	40.7
Fixed	log-Users	-307.9	41.9
Fixed	Fixed	-437.1	46.9

Table 5: ELDP model selection for GDP and number of user analysis, ordered from top (best) to bottom (worst).