Hard Numbers: Language Exclusion in Computational Linguistics and Natural Language Processing

Martin Benjamin
Kamusi Project International
Place de la Gare 12 C
1020 Renens, Switzerland
martin@kamusi.org

Abstract

The intersection between computer science and human language occurs largely for English and a few dozen other languages with strong economic or political support. The supermajority of the world's languages have extremely little digital presence, and little activity that can be forecast to change that status. However, such an assertion has remained impressionistic in the absence of data comparing the attention lavished on elite languages with that given to the rest of the world. This study seeks to give some numbers to the extent to which non-lucrative languages sit at the margins of language technology and computational research. Three datasets are explored that reveal current hiring and research activity at universities and corporations concerned with computational linguistics and natural language processing. The data supports the conclusion that most research activity and career opportunities focus on a few languages, while most languages have little or no current research and little possibility for the professional pursuit of their development.

Keywords: under-resourced languages, computational linguistics, NLP, language technology

1.0. Introduction

This paper looks at technology for "under-resourced" languages by examining the amount of career opportunities and research projects in the field. Two data sets were evaluated to provide hard numbers regarding the proportion of high level research in areas related to computational linguistics. The hypothesis for the research was that high level pursuit of technological development for most of the world's languages is not a widely available career option. This hypothesis was fully supported by the data. Without a significant number of people working in the field, resources for under-resourced languages cannot be developed. The numbers in this study, which indicate where the field will be casting its gaze for years to come, give no cause for optimism that the situation will improve.

People who work on excluded languages know from experience that most of the world's languages remain outside of the technological sphere, but do not have numerical ways to demonstrate the extent of the marginalization. In principle, one could look to the amount of software that is localized in each language, but that would involve getting access to thousands of programs, installing them, and enumerating the available user interface languages – an impossible task that one already knows would reveal almost nil coverage for the supermajority of languages. One could hunt for resources per language, and find a corpus here, a spell-checker there, and a bunch of Wikipedia stub pages about asteroids somewhere else[1]. In the end, though, a spreadsheet with 7000 languages and all known technologies would show a smattering of ticks for a long tail of languages, for example where a passionate developer created an Android app[2] or where a field linguist shared a dictionary on Webonary,[3] and a huge clustering of resources for a small assortment of languages that could be guessed without looking at the data. Kornai's impressive effort (2013) to quantify existing resources for languages of the world found that 6,541 had no detectable live online presence. Following Scannell (2013) and Gibson (2014), it is possible to find instances of usage of nearly 2000 languages within communication technologies such as Twitter, but these are examples of technology as a vessel rather than an avenue for development. Because data about the topic of research activity is not obviously available, we have been left to make impressionistic assertions about the paucity of work in the field.

This study examined three sources of data that provide numerical indications of the extent to which under-resourced languages are active within the overall profession of language technology. The first

[1] The Yoruba Wikipedia, https://yo.wikipedia.org, has more than 31,000 articles listed. However, most of those contain bogus content, including thousands of pages like https://yo.wikipedia.org/wiki/23006_Pazden that are bot-generated stubs containing the names of asteroids. Clicking the link labelled "Ojúewé àrìnàkò" from any Yoruba Wikipedia page gives a random page from the project, with a high probability of landing on an asteroid.

[2] https://mothertongues.org/

[3] https://www.webonary.org/

dataset consists of all the jobs posted on Linguist List (LL) in 2017 that specify applied, computational, or text/ corpus linguistics.[4] The second dataset consists of all the papers and posters presented at COLING 2016,[5] the 26th International Conference on Computational Linguistics, organized by the Association for Natural Language Processing, in Osaka, Japan. The third consists of all the papers and posters presented at ICLDC 2017, 5th International Conference on Language Documentation & Conservation, in Honolulu, Hawaii. None of these are perfect representations of the state of the field, for reasons discussed below, but they give an overall up-to-the-moment indication of the state of attention that under-resourced languages receive among those active in the profession.

Spoiler alert: under-resourced languages receive almost no attention in work related to computational linguistics or natural language processing (NLP).

Type	Number/ Examples	Characteristics
A	4 languages (English, French, German, Spanish)	Massive investment, many existing digital resources, large monolingual and aligned corpora, somewhat functional machine translation with other languages in the group, primary focus of language technology research and development.[6]
B	About 25 languages (many official languages of the EU, Chinese, Japanese, Russian, Arabic)	Moderate to large investment and research, increasing digital resources, large monolingual corpora with bilingual alignment to A languages (especially English), rough machine translation to A languages (usually English), focus of interest for EU and national funders
C	All the rest. Almost 7000 languages, spoken by the majority of the world's 7 billion people.	Zero to mediocre investment and research. Some languages like Swahili and major languages of India, with more than 100 million speakers, have active research communities and rough machine translation, usually to English. A couple of thousand have some form of print dictionary, ranging from lists of a few hundred words to massive volumes with hundreds of pages. Most are 'embattled' – either close to extinction, or disfavored by policy or practice. Funding is usually sparse.

Table 1: Language Categories

2. How much less resourced are “less-resourced” languages?

Table 1 proposes a typology of languages, wherein languages in Category A are the ones that receive high attention in NLP research, Category B languages receive moderate attention, and Category C languages receive little or no attention. Figure 1 shows those languages at exact scale, proportional to the total number of languages in each category: a square with 4 units for Category A, 25 units for Category B, and nearly 7000 units for Category C. Figure 2 was a speculation drawn about two years prior to the present study that posited the ratio of research invested in each category as a rough inverse of the number of languages affected. This paper examines the hypothesis implicit in Figure 2. The hard numbers in this study show that the scale shown for research activity between categories A and C is about right. The representation underestimates the level of activity for Category B languages, however – mid-resourced languages, where Kilgarriff and Grefenstette’s 2003 observations about trends toward increasing digital multilingualism hold true, should have a bigger box, with a gradient toward “under-resourced” that is certainly reached around 50. On the other hand, the data bears out that several languages that were cast as Category B – notably Chinese, Japanese, and Arabic – are benefiting from significant professional attention, and could now be on the borderline of Category A, which would maintain the ratio between A and B closer to the initial depiction. While a discussion of the state of Category B languages is outside of the scope of this paper, as they enjoy a host of advantages that elevate them above any notion of “under-resourced”, it is worth noting that many are making strides that will redound increasingly to their benefit in the years to come.

[4] Records were laboriously reviewed by setting search criteria on the Linguist List jobs page, https://linguistlist.org/jobs/search-job1.cfm. Linguist List keeps records dating back many more years, but procuring those in a practical format would require imposing on their staff. Whether a single year’s data is completely representative is therefore an open question.

[5] http://coling2016.anlp.jp/

[6] The list of “Any language” treated by DeepL (English, German, French, Spanish, Italian, Dutch, and Polish) at http://www.deepl.com/translator could be a better estimate, but discussion of levels of inclusiveness among better-resourced languages would be the subject of a different paper.

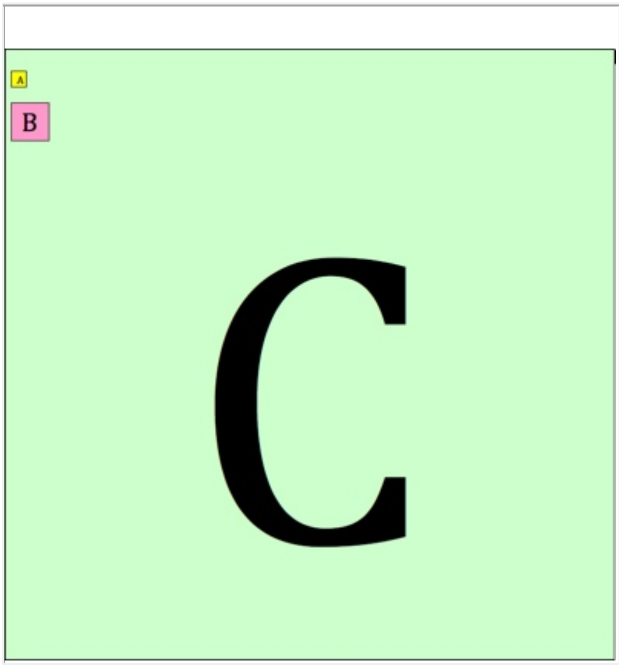	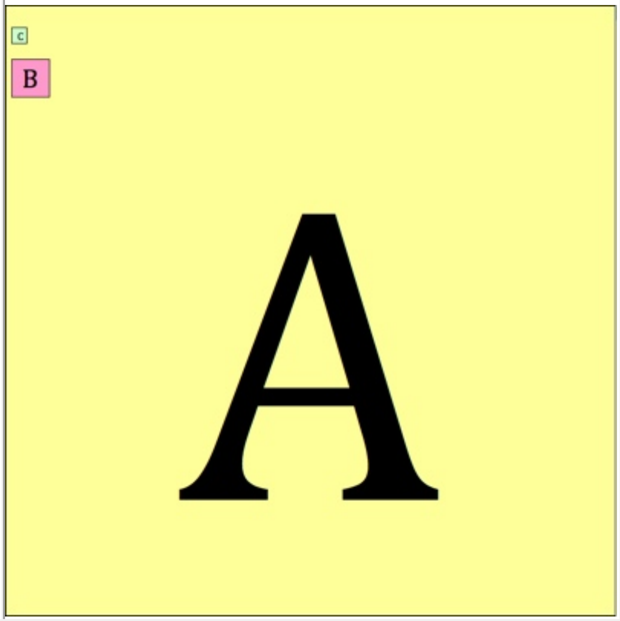
Figure 1: Languages of each type as a proportion of world total (actual numbers)	Figure 2: Languages of each type as a proportion of investment and research attention (hypothesized estimate)

2.1 Linguist List

The LL data shows all 426 jobs posted for applied, computational, or text/corpus linguistics during 2017. Of these listings, 309 mention one or more languages. A total of 42 languages are mentioned, in addition to the categories “African”, “Aboriginal”, “Foreign languages”, “Germanic”, “Indigenous languages of North America”, “Multilingual”, “Pacific Pidgins and Creoles”, “Romance”, and “Turkic”. By far the most frequent language mentioned is English, with 128 listings. Second place goes to 117 unspecified listings. Random inspection shows that “unspecified” often means English, but if not English will almost certainly involve one of the other languages in Category A or B; for example, a position[7] is open to “any language, preferably the languages taught at the Center”: Arabic, Chinese, French, German, Hindi, Italian, Japanese, Korean, Portuguese, Russian, and Spanish. The combination of the four original Category A languages and “unspecified” yields 344 total mentions (with some jobs mentioning more than one language). The next 25 languages or clusters are mentioned 160 times. Finally, 21 languages or clusters are mentioned one time each.

Many of the languages at or near the bottom appear only in jobs announced by the translation company Lionbridge or a company called Pygmalion that is working on NLP. These include several languages of the Indian sub-continent, Tagalog, Thai, and Vietnamese, each spoken by tens of millions of people. The inclusion of these languages may indicate a glimmer of recognition that some excluded languages may harbor a hidden profit potential. However, impressionistic evidence, including professional visits by the author to language technology offices in India and Vietnam, communications with linguists and technologists throughout the region, and interviews with internship candidates for NLP positions from universities around Asia, does not indicate that any of these languages other than Hindi could be considered a candidate for current or imminent inclusion in Category B.

Several languages at the bottom share the profile of most of those that do not even appear: economically and politically powerless, and not considered as candidates for technology by either producers or consumers. These include the Bajau language of Indonesia, Crow in the US, Inuktitut in Canada, Keres in Mexico, the Tok Pisin pidgin of Papua New Guinea, and the African, North American “indigenous” or “aboriginal” languages, and Turkic language clusters writ large. Closer inspection of the announcements reveals that none of these jobs involve NLP. For example, the position for African and Turkic languages[8] is for an undoubtedly fascinating project called “Discourse reporting in African storytelling” for which post docs are expected “to conduct fieldwork collecting traditional narratives, develop an annotated corpus of narrative texts, analyze selected aspects of these texts, and collaborate with other members of the team on theoretical issues related to the encoding of

[7] https://linguistlist.org/issues/28/28-5377.html

[8] https://linguistlist.org/issues/28/28-5291.html

reported discourse". Crow and Keres[9] entailed "cutting and labeling audio, data entry from handwritten notes, additional tasks relating to analysis and organization of the data, and some retyping of existing corpus." No university or corporation on the planet took advantage of the free services of LL to advertise for a linguist to work on the development of a single language of Africa or South America, nor any but the most lucrative or politically well-placed languages of Asia or Europe.

While the LL data is indicative of the global state of hiring, it should not be considered definitive for several reasons. The list only includes jobs where HR or the search committee is aware of the LL job board and considers it important. Posters include universities, translation companies, and some big technology companies such as Amazon and Google. However, many companies that seek employees for NLP elsewhere, such as Angel List,[10] are absent from LL, perhaps because they are more interested in hiring people with a computer science background than with training specifically in linguistics. Further, LL does not penetrate to many national job markets for Category B languages where conferences such as COLING demonstrate that active research is underway, such as Polish, Catalan, and Turkish. We cannot, therefore, make universal claims from the data, but we can use it as strong support for what were previously anecdotal inferences. In particular, the data is not granular enough to support conclusions about which languages belong in Category B or how extensively work is available in those languages. However, matching the LL data against the list of languages that have been identified with ISO 639-3[11] codes shows nearly 7000 zeros: it is beyond doubt that no jobs were available anywhere in 2017 for work on language technology for the supermajority of the world's languages.

2.2. COLING 2016

The COLING 2016 program listed 230 papers and posters. Of these, 16 languages were mentioned by name (English, Arabic, Chinese, German, Hebrew, Hindi, Japanese, Korean, Manipuri, Mongolian, Polish, Sanskrit, Spanish, Thai, Turkish, and Urdu), 131 did not specify a language, 53 papers were classified as "multilingual", and 10 papers were classified as "under-resourced" due to their inclusion in a special track for the topic. Only 13 specified English, but random scanning showed that as the language of analysis for many "unspecified" papers. Many of the papers in the "multilingual" category dealt with machine translation, which is inherently about more than one language, and closer inspection shows to pertain most often to Category A or Category A+B languages. For the many papers that did not specify a language, random inspection showed only Category A or B as languages of concern.

An attempt was made to estimate the language of the research based on the last name of the lead author. No bankable results were achieved, because many names could not even begin to be associated with a language, and many names that indicate the ancestry of a researcher do not indicate their current location, research interests, or available datasets. However, it is probably not a coincidence that the location of the conference in Asia, and the high participation of researchers from Chinese and Japanese institutions, coincided with 94 submissions from people with names associated with China and Japan. Although their paper titles might not have specified Chinese or Japanese as the languages of research, 15 could be identified from their descriptions as pertaining to those languages. Many papers submitted from Asian institutions that focused on deep computational issues, though, such as "Asynchronous Parallel Learning for Neural Networks and Structured Models with Dense Features", often used English as their data core, since that is where they could benefit from and measure themselves against other research; unfortunately, author's institution was only available in the processor-crushing 3500 page proceedings PDF, so the potentially fruitful inquiry of the extent to which English is central to research interests in non-English countries was not practical. In no case did inspection of an article in the proceedings that was not labeled "under-resourced" reveal research in a Category C language, and no papers were submitted to the conference by authors with a name that was discernably African or from an otherwise under-represented language area. Though more detailed research about whether NLP researchers focus on their own languages or the languages with high industry demand could reveal interesting sociological patterns, the present findings about surnames are reported in the spirit of "negative results", a hypothesis tested and found to be unsupported by the evidence at hand.

As with the LL data, COLING data was not extensive enough to make statistically valid claims about the global distribution of research on any given language. Even more than LL, many Category B languages were not represented at all at the conference, though we know that research on languages such as Danish, Dutch, and Romanian is occurring at institutions in the countries where those languages are spoken. However, as with the LL data, we can make certifiable observations about where

[9] https://linguistlist.org/issues/28/28-3570.html

[10] https://angel.co/

[11] https://en.wikipedia.org/wiki/ISO_639-3

research is generally not happening: almost all of the languages in Category C.

2.3. ICLDC 2017

ICLDC[12] is a biennial conference that attracts people working on excluded languages, especially those spoken in countries around the Pacific Rim. The conference program[13] was examined for a reverse perspective on the other two data sets. The question was, among scholars and practitioners of under-resourced languages, what proportion of research activity is given to developing technological resources?

166 papers and posters are listed in the conference program. Workshops and roundtables were not considered. The titles were judged on the single criterion of whether they pertained in a broad way to digital technology. "Creating a Digital Shell for Indigenous Language and Culture Sharing" was considered relevant, whereas "Languages, 'Languoids', and ISO-codes for Language Diversity and Variation" was judged to be outside the scope of technology development. The assumption was that all papers dealt with Category C languages. About 75 languages are indexed in the conference program, with African languages having very little representation.

Twenty-four papers, or 14.5%, met the criterion for relevance to improving technological resources. Most of these were discussions of the creation of particular data resources or learning tools. For example, "Leveraging Web Technologies to Enrich Archival Materials for Use in Language Revitalization" is a discussion of the digitalization and use of archival materials for an Alaskan language; such a corpus building activity is foundational for potential future NLP, but does not involve computational advances *per se*. Similarly, "Large-scale Language Documentation in Nepal: A strategy based on SayMore and BOLD" is about the use of software to produce data, not the development of software itself. "Re Tlli7sa ell re uqw7úq̓wis: Engaging Indigenous language learners with an epic story through a language learning app", an example of how technology can be used in the service of endangered languages, is about the use of digital tools, not their production.

The ICLDC data demonstrate that work on language technology and work on under-resourced languages are conducted by almost completely different groups of people. This is correlated to the jobs board on LL, where some (not many) positions regarding under-resourced languages can be found using "language documentation" and "lexicography" as the search criteria instead of the technology-oriented criteria stated in Section 2.1. Similarly, the fourteen chapters of Day, Rewi, and Higgins (2016) that deal with contemporary research on "besieged" languages give only glancing mention to possible inclusion on the Internet or within communication technologies. Succinctly: research activity on non-lucrative languages rarely intersects with the development of language technology resources.

2.3.1. Coda: ComputEL-2

As a counterpoint to the previous sections, mention should be made of a specific effort to assemble practitioners of technology for under-resourced languages. ComputEL 2 was the second Workshop on Computational Methods for Endangered Languages, [14] held immediately after ICLDC 2017. This workshop featured 23 presentations[15] on themes related to excluded languages and technology. Many of these had concerns similar to those of ICLDC, developing and using digital data, such as "Endangered Data for Endangered Languages: Digitizing Print dictionaries". However, a few of the papers could have been presented at COLING instead. For example, "Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network" and "Converting a comprehensive lexical database into a computational model: The case of East Cree verb inflection" both deal with the sorts of issues that are at the forefront of computational linguistics. Given that under-resourced languages such as East Cree and Inuktitut are, in aggregate, spoken by more than half the world's population, it should be shocking that one has to scour the planet for a small workshop devoted to their advances within language technology, not just as a question of equity but as one of research opportunity.

3.0. Conclusions

As seen at ComputEL, excluded languages are viable candidates for the computer science aspects of language research. In fact, one could argue that a topic such as East Cree verb inflection provides a challenge that is more likely to push the edges of computational linguistics than yet another foray into English data. Thousands of fascinating research questions that could push the frontiers of NLP are not being asked, because technology research is trapped in a small set of well-picked-over languages,

[12] http://icldc5.icldc-hawaii.org/

[13] http://icldc5.weebly.com/uploads/2/4/9/6/24963413/icldc_5_program.pdf

[14] http://altlab.artsrn.ualberta.ca/computel-2/

[15] http://altlab.artsrn.ualberta.ca/computel-2/computel-2-accepted-presentations/

while the limited opportunities for under-resourced languages do not involve pursuing their digital futures. From an intellectual perspective, the chasm between computation and Category C languages represents many lost opportunities for scientists to forge into fresh, uncluttered territory. For corporations, blinders about the profit potential of diverse languages could be overlooking vast markets, particularly for about 350 languages with more than a million speakers but virtually no technological presence. To give one final data set, examine the "List of LREC 2016 Shared LRs" (Language Resources);[16] or, on a blind bet, the forthcoming list for 2018; scrolling through the list, there is no need to get an exact count to see the extent to which LREC members produce resources almost exclusively for Category A and B languages. Though the datasets are too small to draw iron-clad statistical conclusions, the data evaluated in this paper gives hard numbers beneath a hard truth: not only are most languages currently neglected from the digital sphere, but today's hiring and research activity destine that exclusion to continue without end. Computational linguistics and NLP hardly intersect with the supermajority of the world's languages, job positions rarely appear to pursue such intersections, and research for most languages remains perpetually stalled.

References

Day, D., Rewi, P., and Higgins, R., eds. (2016) The Journeys of Besieged Languages. Cambridge Scholars Publishers.

Gibson, M. (2014). A framework for measuring the presence of minority languages"in cyberspace. Presentation to the 3rd International Conference on Linguistic and Cultural Diversity in Cyberspace, Yakutsk, Russia.

Kilgarriff, A., and Grefenstette, G. (2003) Introduction to the special issue on the web as corpus. Journal of Computational Linguistics – Special issue on web as corpus. Volume 29, Issue 3, September 2003, Pages 333-347.

Kornai A (2013) Digital Language Death. PLoS ONE 8(10): e77056. doi:10.1371/journal.pone.0077056.

Scannell, K. (2013). How many languages are on the web? The Crúbadán project 10+ years on, invited talk at the Workshop on Corpus-based Quantitative Typology (CoQuaT 2013), Leipzig, 14 August 2013

[16] http://lrec2016.lrec-conf.org/en/shared-lrs/