NLP for Language Documentation: Two Reasons for the Gap between Theory and Practice

Luke Gessler **Katharina von der Wense**
University of Colorado Boulder
{luke.gessler,katharina.kann}@colorado.edu

Abstract

Both NLP researchers and linguists have expressed a desire to use language technologies in language documentation, but most documentary work still proceeds without them, presenting a lost opportunity to hasten the preservation of the world's endangered languages, such as those spoken in Latin America. In this work, we empirically measure two factors that have previously been identified as explanations of this low utilization: curricular offerings in graduate programs, and rates of interdisciplinary collaboration in publications related to NLP in language documentation. Our findings verify the claim that interdisciplinary training and collaborations are scarce and support the view that interdisciplinary curricular offerings facilitate interdisciplinary collaborations.

1 Introduction

In 2019, 5 out of 68 indigenous languages from Colombia were about to become extinct: one of them, Tinigua, had only a single speaker left;[1] for the others, the situation looked only marginally better. Globally, approximately half of humanity's roughly 7,000 languages are considered endangered (Bromham et al., 2022). While many people in Latin America and other places around the world want their languages to be preserved, language documentation – the process of producing grammars and texts to record a language – is very labor-intensive. Demand for individuals who can perform language documentation far outstrips supply worldwide, and there is little reason to expect this will change any time soon.

In the past 20 years,[2] the computational linguistics (CL) and natural language processing (NLP) communities have responded with systems which can automate some of the labor required in language documentation (LD). For example, ELPIS (Foley et al., 2018) can transcribe audio into text even under the challenging conditions endemic to the LD process, such as low data volumes. Despite the considerable number of computational systems which have been proposed and described over this period, they have seen little practical use (see, e.g., Good et al. 2014; Flavelle and Lachler 2023).

It is puzzling, *prima facie*, that systems with proven potential to facilitate LD have not been integrated into LD projects, and several explanations of this have been offered: Gessler (2022) cites lack of interoperability between NLP systems and LD apps. Flavelle and Lachler (2023) cite an array of organizational barriers that linguists, NLP researchers, and community members face in their collaborations, including conflicting professional incentives and a lack of understanding of the other party's conceptual frameworks. They further observe that "coursework in computational linguistics is rarely required (or even available) to students training to be documentary linguists, and vice-versa", with the consequence that they "miss out on the opportunity to learn even the basic concepts of each other's fields, they also miss out on the opportunity to build connections with others who may go on to specialize in those areas". We expect that there is plenty of room for many explanations to be correct, as this issue is multifaceted.

In this work, we aim to quantify two potential reasons for the lack of usage of NLP systems in real-world LD projects and to compare situations across countries. Specifically, we ask the following research questions (RQs): (1a) How many top-25 universities offer graduate programs in which students can learn about both NLP and LD? (1b) How does the answer to the aforementioned question differ across countries? (2a) What percentage of papers on NLP for LD are the result of truly interdisciplinary collaborations between NLP re-

[1] https://www.eafit.edu.co/noticias/eleafitense/113/universo-linguistico-colombiano-68-idiomas-propios

[2] For an early example, see Kuhn and Mateo-Toledo (2004).

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 1–6

searchers and documentary linguists? (2b) How does the answer to the aforementioned question differ across countries? (3) Finally, is there a connection between the answers to (1a) and (2a)?

To answer our RQs, we use publicly available data from two sources: graduate program curricula and academic publications. We treat each *country* as as an individual unit: quantities we gather are aggregated per country before we proceed with analysis. Aggregation at any smaller unit (e.g., at the university or individual level) would make data collection impractical, and while it is true that countries are not monolithic with respect to curricular offerings or publishing cultures, we observe that these differences are in sum much more pronounced between rather than within countries.

2 University Curricula

We examine five countries: the United States, Germany, Brazil, Mexico, and Colombia. We choose the United States and Germany because of their prevalence in AI publication venues and because their academic cultures are quite distinct. We additionally choose Brazil, Mexico, and Colombia, as these countries, like much of Latin America, have many indigenous languages.

For each country, we consider the 25 top-ranked universities according to QS World University Rankings 2024.[3] For each university, we then determine whether it offers a graduate program in computer science (CS) or linguistics (Ling). We define a "graduate program" as anything that is at least partially beyond the scope of a United States bachelor's degree: any MS or PhD program would qualify, though some degree programs such as *licenciaturas* vary in whether they include graduate-level training, and we examine their curricula on an individual basis.

As for whether a program qualifies as "computer science" or "linguistics", we would like to capture the programs that have the highest densities of NLP researchers and documentary linguists. To this end, we define a "computer science" program as any program that has "natural language processing" or "computational linguistics" in its name, or has a graduate course in algorithms; and we define a "linguistics" program as any program that offers at least one graduate course in theoretical linguistics.[4]

[3] https://www.topuniversities.com/world-university-rankings

[4] We do not consider whether the program contains the word "linguistics", as this would include programs that are primarily focused on language teaching or learning and would be unlikely to host a documentary linguist student.

Country	CS Dept	Ling Dept	Both Depts	NLP in CS, LD in Ling	LD in CS	NLP in Ling	NLP & LD in Ling
USA	25	17	17	12	0	12	9
Germany	25	20	20	1	0	6	0
Mexico	12	3	2	0	0	1	0
Brazil	23	12	12	1	0	2	1
Colombia	8	4	3	0	0	1	0

Table 1: University curriculum data by country. Among the 25 top-ranked universities in each country, the columns display the number of universities which have a qualifying computer science program; have a qualifying linguistics program; have both programs; have both programs *and* offer both an NLP course in the computer science program and a LD course in the linguistics program; have an LD course in the computer science program; have an NLP/CL course in the linguistics program; and have both an NLP/CL and an LD course in the linguistics program.

For each eligible program, we determine whether it offers coursework in NLP/CL or LD. For a CS department, an NLP course must be dedicated to just NLP (an introduction to AI with a couple of weeks introducing NLP does not qualify) and a LD course must either include real LD or study systems which are explicitly intended for use in LD settings. For a linguistics department, an NLP/CL course should cover the use and/or development of NLP systems which can automatically perform linguistic analysis (such as finite-state automata, PoS taggers, or parsers) or modern NLP, and a LD course should be structured as a typical field methods course where students document a language through the full term of the course.

2.1 Results

We give a summary of our findings in Table 1. See [REDACTED] for full data.

RQ1a First, for CS departments, we can see that none of the 93 departments offered an LD-related course. Of the 56 linguistics departments, 22 offered an NLP/CL course, and 10 offered both an NLP/CL course and an LD course. This indicates low overall availability of interdisciplinary training to both populations, as it is overall not common for graduate students to take courses outside of their

departments.

RQ1b Considering now the differences between countries, we can see that 9 out of 10 of the universities offering both an LD and NLP/CL course are in the US, as are 12 out of the 22 offering an NLP/CL course. We also see that 12 out of the 14 universities which have both a CS program with an NLP course and a linguistics program with an LD course are in the US, which we view as potentially facilitative of interdisciplinary collaborations. Thus while rates of interdisciplinary training and contact are low overall, they are comparatively higher for the US, and if the view that this ought to encourage collaboration is correct, then we should expect to see higher rates of interdisciplinary publications among works from the US (cf. RQ3).

3 Publications

We collect a large number of publications, each with the following annotations:

1. Relevance – whether the work's core contribution is a resource or system that *could* directly aid the efforts of documentation projects. We operationalize this requirement in two ways. First, a relevant paper must use at least one dataset that is an order of magnitude smaller (by tokens/hours) or more than typical high-resource datasets for the task, and the language of this dataset must be unrepresented among these high-resource datasets. (For example, a work on Universal Dependencies parsing that used the Thai treebank would count, because at 22K tokens, the Thai treebank is over an order magnitude smaller than a typical English treebank, EWT, which has 250K tokens.) We further require that a relevant paper's task be one that has direct relevance to LD activities, such as morphological parsing or machine translation.

2. Country – the country with the most representation among the authors' organizational affiliations. (We do not consider authors' nationalities, whatever they may be—only their institutional affiliations at the time of the work's publication.) If there is a tie, we take the country of the first author.

3. Documentation as Purpose (DaP) – whether LD was explicitly mentioned in the paper as a motivation for the work.

4. Performance of Documentation (PoD) – whether the collection of novel documentary data was a part of the work, where "collection" means the creation of digital primary language data that did not exist before the work.

5. Interdisciplinarity (Int) – whether the author list contains at least one NLP researcher and one documentary linguist. Any individual researcher may belong to at most one of these groups. Authors are assessed on the basis of what venues they have published in: typical NLP venues include ACL conferences, and typical documentary linguistics venues include LD&C and ICLDC.

The population of relevant (as defined above) papers is diverse and distributed throughout many publication venues, which makes it non-trivial to sample from it. We employ two resources for gathering data which have complementary strengths: the ACL Anthology,[5] a machine-readable repository of publications from venues associated with the Association for Computational Linguistics, and Semantic Scholar,[6] an academic publication aggregator with advanced querying capabilities.

AmericasNLP & ComputEL The first part of our data comes from two venues contained in the ACL Anthology which we identify as having the highest potential density of relevant papers of any publication outlet we are aware of. These are the AmericasNLP workshop[7] and the ComputEL workshop.[8] All documents which belong to one of our target countries are annotated. This dataset is useful because of its concentration of highly relevant papers, but its weakness is that it is biased heavily towards the relevant papers that are most concerned with LD as a primary goal.

Semantic Scholar The second part of our data comes from Semantic Scholar's bulk search feature, which we use to find documents which contain at least one keyword related to LD and at least one keyword related to NLP.[9] Results are shuffled, and the first 50 relevant papers for each of our target countries are annotated. This dataset is useful because it ought to offer a wider view of relevant papers, but its weakness is that its keyword-based approach likely excludes many relevant papers.

3.1 Results

For any one of the Latin American countries we consider in the previous section, we are unable to find more than 5 relevant papers despite an exhaustive review of the over 3,000 publications that were

[5] https://aclanthology.org/
[6] https://www.semanticscholar.org/
[7] https://aclanthology.org/venues/americasnlp/
[8] https://aclanthology.org/venues/computel/
[9] See Appendix A for details.

Country	DaP	PoD	Int	Total
ComputEL & AmericasNLP				
USA	37	11	22	48
Germany	3	1	1	5
Total	*40*	*12*	*23*	*53*
Semantic Scholar				
USA	18	6	8	50
Germany	4	3	5	50
Total	*22*	*9*	*13*	*100*

Table 2: The number of relevant papers for each country (Total) which respectively had documentation as an explicit purpose (DaP), actually performed documentation (PoD), and had an interdisciplinary authorship (Int).

returned by our query, and we therefore consider only Germany and the United States in this section. We give a summary in Table 2, and all data is publicly available at [REDACTED].

RQ2a Looking at the *Total* rows in Table 2, we see that less than half of all relevant papers are the results of truly interdisciplinary collaborations – for Semantic Scholar, as little as 13%. While not all papers that could be relevant for LD necessarily benefit from being interdisciplinary, we claim that this is desirable at least for papers that cite LD as their main motivation. As those papers number 62 overall, while only 36 are interdisciplinary, we find the latter number to be unfortunately small. This shows that there is much room for growth in the formation of interdisciplinary collaborations.

RQ2b For the ACL Anthology data, the United States has significantly more representation than Germany, and around 80% of works name LD as an explicit goal. Curiously, a large but smaller number of works are interdisciplinary, which could be interpreted as evidence of a degree of awareness within the NLP community in the United States of the need for NLP in LD.

A different but consistent picture emerges in the S2 data. While many American publications still cite documentation as a motivation, the proportion is smaller, and the number of interdisciplinary authorships is also smaller. This corroborates our initial conjecture that ComputEL and AmericasNLP papers would be disproportionately focused on documentation relative to the population of relevant papers as a whole. Fewer than 10% of German papers cite documentation as a purpose or have an interdisciplinary team.

RQ3 Unfortunately, the amount of papers we are able to find, especially for Latin American countries, is too small to give a definite answer to RQ3. However, the fact that more US-based than Germany-based researchers motivate their work with LD and the larger number of interdisciplinary paper collaborations could be interpreted as evidence of a higher degree of awareness of LD challenges in the NLP community as well as a larger number of LD researchers who are aware of NLP. This, in turn, could potentially stem from more readily accessible education on LD as well as from programs that offer courses in both LD and NLP.

4 Conclusion

We have presented what is to our knowledge the first evidence that provides an empirical understanding of two factors in the adoption of language technologies in LD: university curricula and collaboration trends between NLP researchers and documentary linguists. Our data confirms previous claims that rates of interdisciplinary training and collaboration are low, even for work that cites application in language documentation as a motivation.

Moreover, while the scale of our data precludes a firm conclusion, it is consistent with the claim that interdisciplinary coursework is a partial determinant of collaboration rates, as the higher rates of interdisciplinary course offerings in the United States (relative to Germany) are mirrored by higher rates of interdisciplinary publishing by authors working in the United States. This broadly supports the view that interdisciplinary graduate coursework is important for supporting the incorporation of human language technologies into LD practice. More evidence is needed, however, in order to investigate other possible factors: perhaps other influences, such as nation-level grant programs or academic cultures, are directly affecting both curricula and rates of interdisciplinary collaborations.

We therefore join Flavelle and Lachler (2023) in identifying interdisciplinary curricular offerings as an important way for the NLP and linguistics communities to work towards the ultimate goal of aiding LD with language technologies. Additionally, we observe that many of the same benefits could be gained from interdisciplinary workshops such as the LTLDR workshop (Neubig et al., 2020), which gathered documentary linguists, NLP researchers, and community members for the explicit purpose of fostering interdisciplinary collaborations.

Limitations

Our findings are limited by the quantity of data that we have collected and the methods we used to sample the data points that we have. For universities, this comes out in our selection of 5 particular countries and our consideration of 25 universities from each, as we were unable to include more countries and universities given the high time cost of annotating a single university. For publication data, this is instantiated in our two methods for collecting papers which, as we described, we expect introduced sampling bias, though these two methods seemed that they would introduce the least sampling bias of any of the other methods we considered while still remaining practical to perform.

References

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Jonas Kuhn and B'alam Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma, and Patrick Littell. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

A Search Criteria

We use Semantic Scholar's bulk search API[10], which accepts queries in a rich structured format which features several operators which form trees over keyword arguments. Our query is provided below in an abstract syntax tree. The two main parts of it contain keywords related to language documentation and NLP models, respectively. The tilde operator `x~n` specifies that up to `n` words may intervene between the words in `x`. Both keyword lists are joined with the logical or operator `|` which is satisfied if any one of the keyword expressions are satisfied, and both keyword lists are finally joined with the logical and operator `+` which is satisfied only if both subexpressions are satisfied.

```
[
    "+",
    [
        '|',
        '"low-resource"',
        '"low resource"~1',
        '"less-resourced"',
        '"less resourced"~1',
        '"under-resourced"',
        '"under resourced"~1',
        '"under-studied"',
        '"under studied"~1',
        '"less-studied"',
```

[10] https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_bulk_search

```
        '"less studied"~1',
        '"endangered language"~1',
        '"indigenous language"~1',
        '"language documentation"',
        '"document language"',
        '"language revitalization"',
        '"revitalize language"',
        '"language maintenance"',
        '"maintain language"',
        '"language revival"',
        '"revive language"',
        '"ELAN"',
        '"FLEx"',
        '"FieldWorks Language Explorer"',
        '"LingSync"',
        'typological',
    ],
    [
        '|',
        'model',
        'resource',
        'lexicon',
        'parser',
        'corpus',
        'dataset',
        'document',
        'dictionary',
        'grammar',
        'segmentation',
        'orthographic',
        'normalization',
        'evaluation',
        'experiments',
        '"machine translation"',
        '"automatic translation"',
        'predict',
        'neural'
    ]
]
```