Endangered Languages meet Modern NLP

Antonios Anastasopoulos[†] **Christopher Cox**[◇] **Hilaria Cruz**[♣] **Graham Neubig**[‡]
[†]Department of Computer Science, George Mason University
[‡]Language Technologies Institute, Carnegie Mellon University
[◇]School of Linguistics and Language Studies, Carleton University
[♣]Comparative Humanities, University of Louisville
antonis@gmu.edu gneubig@cs.cmu.edu
christopher.cox@carleton.ca hilaria.cruz@louisville.edu

1 Description

Computational Linguistics and Natural Language Processing (NLP) have taken immense strides, spearheaded by neural methods and large data collections. The result is ubiquitous language technology and vast amounts of research on new tasks and products. However, the vast majority of the world's languages have been mostly ignored, including the most vulnerable among them: endangered languages.

The lack of communication between the NLP community and the documentary linguistics community is partly to blame (Bird, 2009). Even though field and documentary linguists produce resources and use NLP methods, this is done in isolation, as computational methods are seen as a means towards the final goal, which typically is language description. The extreme pace of language loss and the urgent needs for language revitalization, however, require that we utilize documentations and go beyond language description: enter 21st century NLP.

Himmelmann's 20-year-old radical vision (Himmelmann, 1998) for a data-centric approach to language documentation (which sparked the creation of modern documentary linguistics) has slowly begun to materialize (McDonnell et al., 2018). For example, the Workshops on the Use of Computational Methods in the Study of Endangered Languages (Comput-EL) (Good et al., 2014; Arppe et al., 2017; Arppe et al., 2019) have provided a small forum for the much-needed discussion between NLP practitioners and documentary and field linguists.

Meanwhile, increasingly more focus is dedicated on NLP research and bringing modern technologies to endangered languages. For example, mobile applications have been developed for data collection (Bird et al., 2014; Gauthier et al., 2016) and are actively used in documentation projects (Blachon et al., 2016); automatic speech recognition models have been created to aid with automatic phonetic or orthographic transcriptions focusing in indigenous Australian (Foley et al., 2018) or tonal languages from China and the Americas (Michaud et al., 2018); machine translation for under-represented languages have been presented as new corpora have been collected (Abbott and Martinus, 2018; Abate et al., 2018); cross-lingual transfer has been successfully applied for tagging, morphological analysis and inflection (McCarthy et al., 2019; Anastasopoulos and Neubig, 2019); multitask and active learning are being used for learning from continuous annotations on multiple tasks (Gerstenberger et al., 2017; Anastasopoulos et al., 2018; Chaudhary et al., 2019); approaches dedicated to indigenous polysynthetic languages have been developed (Schwartz et al., 2019; Kann et al., 2018); and computational methods have been used to study or discover typological features from large collections of text (Asgari and Schütze, 2017; Malaviya et al., 2017).

In this tutorial, we will outline the language documentation process and revitalization efforts, while also mapping them to concrete computational tasks. We will then focus on the machine learning approaches tailored to tackle these tasks under this very data-constrained setting. An overview of many of those NLP methods, as applied for language documentation, can be found in co-proposer Anastasopoulos' PhD dissertation (Anastasopoulos, 2019). Other surveys focus on the state of language technologies within specific geographic areas, such as co-proposer Cox's overview of Canadian languages (Littell et al., 2018) or the one by Mager et al. (2018), focusing on indigenous American languages.

The goal of our tutorial will be two-fold. On one hand, we will aim to acquaint the audience with the needs of the documentary linguistics community, and cover the already existing computational research

Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts, pages 39–45
Barcelona, Spain (Online), December 12th, 2020.

in the field. On the other hand, we will discuss the capabilities and limitations of current computational approaches, so that the participants will know when and how to apply NLP methods, as well as how they could collect data and create corpora that can be used by NLP methods to aid both documentation and computational work. Ideally, by the end of our tutorial, the attendees will be familiar with the current research and the state-of-the-art in NLP for endangered language documentation and revitalization *and* be aware of the many standing challenges that lie ahead.

First, we will introduce the challenges posed by language documentation and revitalization, such as the transcription bottleneck, and how machine learning methods can fit into the pipeline. Unlike what is considered a typical NLP setting, working with endangered language data has intricate nuances: the lack of standard orthography, or even complete absence of a writing system; the extremely limited amount of data; language typologies widely different than anything used/tested in prior work; and even the lack of established benchmarks. We will discuss these nuances in depth, and how they relate or can be remedied with existing NLP research. We will also provide example code for many of these methods, and show how standard NLP pipelines need to be modified in order to account for these nuances. Finally, we will close our tutorial by discussing open problems and challenges.

Relevance to linguistics community This tutorial will bring together two linguistics communities, documentary/field linguists and NLP practitioners. As a result, we hope, the tutorial will *build enough capacity* of computational researchers that will be not only interested in NLP for endangered languages, but also aware of the current approaches and challenges. Elevating endangered languages NLP research is necessary towards bringing these under-represented communities to the spotlight, as speakers of such endangered languages frequently lack the skills to build NLP tools themselves.

Tutorial type The proposed tutorial combines the introductory and cutting-edge tutorial types. The acquaintance of computer scientists with the language documentation process will, by necessity, be at an introductory level. At the same time, though, the tutorial will cover cutting-edge NLP methods and their application to the endangered languages domain.

2 Tutorial Structure

The tutorial will be structured in order to be informative for both linguists (documentary, computational, or otherwise) and for computer scientists who are interested in performing computational work for language documentation and revitalization.

We aim for a three-hour tutorial that covers a reasonable range of all important aspects of this area. Times for the proposed structure are approximate, and they might be adjusted as we refine the tutorial content.

Part 1 [1h 20min]

- Introduction (45 minutes)
  - The language documentation process
  - The transcription bottleneck
  - The NLP tasks that are part of the documentation process
  - The NLP tasks that are necessary for revitalization
- When Data Assumptions Break (35 minutes)
  - The nuances of working with endangered language data
  - Code switching, code mixing, unstable orthographies and analyses, and how to deal with them
  - Creating corpora useful for both computational and linguistics research

Part 2 [1h 40min]

- NLP in extremely low-resource settings (40 minutes)
  - Speech Transcription
  - Tagging and other Annotation
  - Digitization and dictionary/grammar creation
- Multilingual/Polyglot models and cross-lingual transfer (30 minutes)
  - Using and choosing high-resource related languages

  - Multilingual models
  - Leveraging monolingual data and data augmentation
- Working with and addressing the needs of endangered language communities (20 minutes)
  - Community-centered documentation and revitalization
- Conclusion (10 minutes)
  - Resources for linguists
  - Resources for computer scientists

3 Recommended Reading List

The reading list is indicative of the multi-disciplinary coverage of this tutorial. We **highlight** the tutorial presenters in the papers they have co-authored:

1. Reflections on Language Documentation: 20 Years after Himmelmann 1998. Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors. *Language Documentation and Conservation Special Publication No. 15*, University of Hawai'i at Manoa, 2018.

2. Challenges of language technologies for the indigenous languages of the Americas. Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, Ivan Meza. *Proceedings of the 27th International Conference on Computational Linguistics: pp. 55–69*. 2018.

3. Natural Language Processing and linguistic fieldwork. Steven Bird. *Computational Linguistics: ed. 35(3) pp. 469–474*. 2009.

4. Deploying Technology to Save Endangered Languages. **Hilaria Cruz** and Joseph Waring. *preprint*, 2019.

5. Future Directions in Technological Support for Language Documentation. Daan van Esch, Ben Foley, and Nay San. In *Proceedings of the Workshop on Computational Methods for Endangered Languages* (Vol. 1, No. 1, p. 3). 2019.

6. Indigenous language technologies in Canada: Assessment, challenges, and successes. Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, **Christopher Cox**, and Marie-Odile Junker. In *Proceedings of the 27th International Conference on Computational Linguistics: pp. 2620–2632*. 2018.

7. Automatic speech recognition for under-resourced languages: A survey. Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. *Speech Communication 56*: pp. 85–100. 2014.

8. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, **Graham Neubig**, and Séverine Guillaume. *Language Documentation and Conservation*: pp. 481-513. 2018.

9. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). Ben Foley, Joshua T. Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath et al. In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages SLTU: pp. 205-209*. 2018.

10. Pushing the limits of low-resource morphological inflection. **Antonios Anastasopoulos** and **Graham Neubig**. In *Proceedings of the of the 2019 Conference on Empirical Methods in Natural Language Processing*. to appear.

4 Diversity Considerations

This tutorial has been constructed with a focus on encouraging diversity in all aspects:

- The content will aim to encourage diversity in Computational Linguistics and NLP, by promoting and encouraging research on the most under-represented group of languages: extremely low-resource and endangered ones.
- We will use real endangered language data for all examples, ranging from Mesoamerican languages to European dialects to indigenous languages of Asia.
- The instructors' team is fairly diverse both in terms of gender and in terms of seniority (one post-doctoral associate, three assistant professors). The presenters are affiliated with three different institutions from two different countries (from the US and Canada).

5 Prerequisites

- Machine Learning: a basic understanding of modern neural models.
- Programming and other tools: All code examples will be provided in Python, so knowledge of Python and basic command-line tools will be necessary in order to follow along.

6 Tutorial Presenters

Antonios Anastasopoulos is a Post-doctoral Associate at the Language Technologies Institute at Carnegie Mellon University. His interests include various aspects of multilingual Natural Language Processing and Machine Learning, with the main focus being Machine Translation and Speech Recognition for endangered languages and low-resource settings in general. He completed his Computer Science PhD at the University of Notre Dame, with a dissertation on *"Computational Tools for Endangered Language Documentation"*. He has been involved with documentation efforts on Griko, an endangered Greek dialect spoken in south Italy. He co-organized the workshop on Language Technology for Language Documentation and Revitalization, hosted in Pittsburgh in August 2019.
website: `http://www.cs.cmu.edu/~aanastas/`

Christopher Cox is an Assistant Professor in the School of Linguistics and Language Studies at Carleton University. His research centres on issues in language documentation, description, and revitalization, with a special focus on the creation and application of corpora representing Indigenous and minority languages. For the past twenty years, he has been involved with community-based language documentation, education, and revitalization efforts, most extensively in partnership with speakers of Plautdietsch, the traditional language of the Dutch-Russian Mennonites, and with Dene communities in western and northern Canada. He has served as an invited instructor in the area of language documentation and revitalization for community-based, national, and international events, delivering workshops for the Canadian Indigenous Languages and Literacy Development Institute (CILLDI), the Institute on Collaborative Language Research (CoLang), and the American Association for Corpus Linguistics (AACL), among others.
website: `https://carleton.ca/slals/people/cox-christopher/`

Hilaria Cruz is a field linguist with a focus on indigenous languages of Mexico, especially the Chatino languages (Otomanguean), which she speaks natively. She is also part of an interdisciplinary community of linguists and computer scientists who are working to create tools for automatic or semi-automatic transcription and analysis of audio and visual information for endangered languages. They have created a time-aligned speech corpus of transcribed, annotated with parts of speech, and translated Chatino data, which are available on an open-source basis. As a native person and field linguist, she has had many opportunities to teach linguistics in diverse settings and with diverse groups of students. She has taught general linguistics at the university level, in community organizations, and within Chatino communities. Her research interests include ASR for endangered languages, Chatino morphology, and promoting reading and writing in indigenous languages.

Graham Neubig is an assistant professor at Carnegie Mellon University specializing in natural language processing and machine learning. One of his major research interests is methods for low-resource language processing, and specifically for aiding the documentation of endangered languages. He has previously given well-attended tutorials at NLP conferences (EMNLP and NAACL) and the Lisbon and CIFAR Machine Learning Summer Schools. He has won a number of best papers at NLP venues (e.g. EMNLP2016, EACL2017, NAACL2019) and given a number of invited talks on the proposal topic of low-resource language processing, including at Google, UMass Amherst, and New York University.
website: `http://www.phontron.com/`

7 Details

Breadth We aim to provide the first wide coverage overview of NLP approaches for endangered languages. Out of the 25 referenced works (which is not an exhaustive list of the works we will cover) only 7 are co-authored by the presenters, and we expect only about 30% of the tutorial to be based on the presenters' prior work. Similarly, 60% of the suggested reading has not been co-authored by the presenters.

Audience Size Estimate The first three iterations of the Comput-EL workshops (which are very relevant to the theme of our tutorial) have been steadily growing in size, while a regional meet-up organized in Pittsburgh by co-presenters Anastasopoulos and Neubig drew about 40 participants. Similar tutorials on the use of NLP in the International Conference on Language Documentation and Conservation (ICLDC) also draw large crowds. Given this interest, we expect a healthy audience of at least 60 participants. To our knowledge, no similar tutorial has been given before at any NLP conference.

Technical Equipment Internet Access; the participants could bring their laptops in order to follow along with code examples.

Preferred Venues All venues are appropriate for this tutorial, and the presenters do not anticipate major conflicts.

Open Access We agree to the publication of our slides and a video recording of the tutorial. We will additionally make all other materials (complete reading list, software, example data) openly available online.

References

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, et al. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90.

Jade Z Abbott and Laura Martinus. 2018. Towards neural machine translation for African languages. *arXiv preprint arXiv:1811.05467*.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proc. EMNLP*, Hong Kong, November. to appear.

Antonios Anastasopoulos, Marika Lekakou, Josep Quer, Eleni Zimianiti, Justin DeBenedetto, and David Chiang. 2018. Part-of-speech tagging on an endangered language: a parallel Griko-Italian resource. In *Proc. COLING*, pages 2529–2539. Association for Computational Linguistics.

Antonios Anastasopoulos. 2019. *Computational Tools for Endangered Language Documentation*. Ph.D. thesis, University of Notre Dame du Lac, Notre Dame, Indiana, January.

Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, and Lane Schwartz. 2017. Proceedings of the 2nd workshop on the use of computational methods in the study of endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Antti Arppe, Jeff Good, Mans Hulden, Jordan Lachler, Alexis Palmer, Lane Schwartz, and Miikka Silfverberg. 2019. Proceedings of the 3rd workshop on the use of computational methods in the study of endangered languages volume 1. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*.

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark, September. Association for Computational Linguistics.

Steven Bird, Florian R. Hanke, Oliver Adams, and Haejoong Lee. 2014. Aikuma: A mobile app for collaborative language documentation. In *Proc. of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Steven Bird. 2009. Natural Language Processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

David Blachon, Elodie Gauthier, Laurent Besacier, Guy-Noël Kouarata, Martine Adda-Decker, and Annie Rialland. 2016. Parallel speech collection for under-resourced language studies using the LIG-Aikuma mobile device app. In *Proc. SLTU (Spoken Language Technologies for Under-Resourced Languages)*, volume 81.

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Hong Kong, November.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochv'il, 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 *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 200–204.

Elodie Gauthier, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman. 2016. LIG-Aikuma: a mobile app to collect parallel speech for under-resourced language studies. In *Interspeech 2016 (short demo paper)*.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2017. Instant annotations–applying NLP methods to the annotation of spoken language documentation corpora. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pages 25–36.

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, June.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics.

Katharina Kann, Manuel Mager, Ivan Meza-Ruiz, and Hinrich Schütze. 2018. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. *arXiv preprint arXiv:1804.06024*.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.

Chaitanya Malaviya, Graham Neubig, and Patrick Littell. 2017. Learning language representations for typology prediction. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Copenhagen, Denmark, September.

Arya D McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sebastian J Mielke, Jeffrey Heinz, et al. 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.

Bradley McDonnell, Andrea L. Berez-Kroeker, and Gary Holton, editors. 2018. *Reflections on Language Documentation: 20 Years after Himmelmann 1998*. Language Documentation and Conservation Special Publication No. 15, University of Hawai'i at Manoa.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, pages 393–429.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia LR Schreiner. 2019. Bootstrapping a neural morphological analyzer for St. Lawrence Island Yupik from a finite-state transducer. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, page 12.