ComputEL-3

Proceedings of the

3nd Workshop on the Use of Computational Methods in the Study of Endangered Languages

Volume 2 (Extended Abstracts)

February 26–27, 2019
Honolulu, Hawai‘i, USA

Support:

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

UNIVERSITY OF ALBERTA

University at Buffalo

University of Colorado Boulder

ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

UNT

UNIVERSITY OF NORTH TEXAS

ISBN XXX-X-XXXXXX-XX-X

Preface

These proceedings contain the extended abstracts presented at the 3rd Workshop on the Use of Computational Methods in the Study of Endangered languages held in Hawai'i at Mānoa, February 26–27, 2019. As the name implies, this is the third workshop held on the topic—the first meeting was co-located with the ACL main conference in Baltimore, Maryland in 2014 and the second one in 2017 was co-located with the 5th International Conference on Language Documentation and Conservation (ICLDC) at the University of Hawai'i at Mānoa.

The workshop covers a wide range of topics relevant to the study and documentation of endangered languages, ranging from technical papers on working systems and applications, to reports on community activities with supporting computational components.

The purpose of the workshop is to bring together computational researchers, documentary linguists, and people involved with community efforts of language documentation and revitalization to take part in both formal and informal exchanges on how to integrate rapidly evolving language processing methods and tools into efforts of language description, documentation, and revitalization. The organizers are pleased with the range of papers, many of which highlight the importance of interdisciplinary work and interaction between the various communities that the workshop is aimed towards.

We received 34 submissions as papers or extended abstracts. After a thorough review process, 12 of the submissions were selected as papers (35%) which appear in the first volume of the conference procedings. An additional 7 were accepted as extended abstracts which appear in this second volume of the workshop proceedings. The organizing committee would like to thank the program committee for their thoughtful input on the submissions. We are also grateful to the NSF for funding part of the workshop (award #1550905), and the Social Sciences and Humanities Research Council (SSHRC) of Canada for supporting the workshop through their Connections Outreach Grant #611-2016-0207.

ANTTI ARPPE
JEFF GOOD
MANS HULDEN
JORDAN LACHLER
ALEXIS PALMER
LANE SCHWARTZ
MIIKKA SILFVERBERG

Organizing Committee:

Antti Arppe (University of Alberta)
Jeff Good (University at Buffalo)
Mans Hulden (University of Colorado)
Jordan Lachler (University of Alberta)
Alexis Palmer (University of North Texas)
Lane Schwartz (University of Illinois at Urbana-Champaign)
Miikka Silfverberg (University of Helsinki)

Program Committee:

Oliver Adams (Johns Hopkins University)
Antti Arppe (University of Alberta)
Dorothee Beermann (Norwegian University of Science and Technology)
Emily M. Bender (University of Washington)
Martin Benjamin (Kamusi Project International)
Steven Bird (University of Melbourne)
Emily Chen (University of Illinois at Urbana-Champaign)
James Cowell (University of Colorado Boulder)
Christopher Cox (Carleton University)
Robert Forkel (Max Planck Institute for the Science of Human History)
Jeff Good (University at Buffalo)
Michael Wayne Goodman (University of Washington)
Harald Hammarström (Max Planck Institute for Psycholinguistics, Nijmegen)
Mans Hulden (University of Colorado Boulder)
Anna Kazantseva (National Research Council of Canada)
František Kratochvíl (Palacký University)
Jordan Lachler (University of Alberta)
Terry Langendoen (National Science Foundation)
Krister Lindén (University of Helsinki)
Worthy N Martin (University of Virginia)
Michael Maxwell (University of Maryland, CASL)
Steven Moran (University of Zurich)
Graham Neubig (Carnegie Mellon University)
Alexis Palmer (University of North Texas)
Taraka Rama (University of Oslo)
Kevin Scannell (Saint Louis University)
Lane Schwartz (University of Illinois at Urbana-Champaign)
Miikka Silfverberg (University of Helsinki)
Richard Sproat (Google)
Nick Thieberger (University of Melbourne / ARC Centre of Excellence for the Dynamics of Language)

Laura Welcher (The Long Now Foundation)
Menzo Windhouwer (KNAW Humanities Cluster — Digital Infrastructure)

Table of Contents

Conference Program

Tuesday, February 26th, 2019

08:30–09:00 ***Imin Conference Center, Asia Room/Arrival, coffee and chat***

09:00–09:15 ***Opening remarks***

First morning: Tools and processes for language documentation and description, Corpus creation

09:15–09:45 **An Online Platform for Community-Based Language Description and Documentation. Rebecca Everson, Wolf Honoré and Scott Grimm**

09:45–10:15 **Developing without developers: choosing labor-saving tools for language documentation apps. Luke Gessler**

10:15–10:45 **Future Directions in Technological Support for Language Documentation. Daan van Esch, Ben Foley and Nay San**

10:45–11:15 ***Break***

11:15–11:45 *Towards a General-Purpose Linguistic Annotation Backend*
Graham Neubig, Patrick Littell, Chian-Yu Chen, Jean Lee, Zirui Li, Yu-Hsiang Lin and Yuyan Zhang

11:45–12:15 **OCR evaluation tools for the 21st century. Eddie Antonio Santos**

12:15–12:45 *Building a Common Voice Corpus for Laiholh (Hakha Chin)*
Kelly Berkson, Samson Lotven, Peng Hlei Thang, Thomas Thawngza, Zai Sung, James Wamsley, Francis M. Tyers, Kenneth Van Bik, Sandra Kübler, Donald Williamson and Matthew Anderson

12:45–14:15 ***Lunch***

Tuesday, February 26th, 2019 (continued)

First afternoon: Language technologies – lexical and syntactic

14:15–14:45 *Digital Dictionary Development for Torwali, a less-studied language: Process and Challenges*
Inam Ullah

14:45–15:15 Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. Olga Zamaraeva, Kristen Howell and Emily M. Bender

15:15–15:45 *Applying Support Vector Machines to POS tagging of the Ainu Language*
Karol Nowakowski, Michal Ptaszynski, Fumito Masui and Yoshio Momouchi

15:45–16:15 ***Break***

16:15–16:45 Finding Sami Cognates with a Character-Based NMT Approach. Mika Hämäläinen and Jack Rueter

16:45–17:15 Tokenization and disambiguation of potential compounds in North Sámi grammar checking. Linda Wiechetek, Sjur Nørstebø Moshagen and Kevin Brubeck Unhammer

17:15–20:00 ***Dinner***

Wednesday, February 27th, 2019

09:00–09:30 ***Arrival, coffee and chat***

Wednesday, February 27th, 2019 (continued)

Second morning: Use (and reuse) of corpora and collections, Language technologies – speech and morphology

09:30–10:00 *Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead*
Rogier Blokland, Niko Partanen, Michael Rießler and Joshua Wilbur

10:00–10:30 *A software-driven workflow for the reuse of language documentation data in linguistic studies*
Stephan Druskat and Kilu von Prince

10:30–11:00 ***Break***

11:00-11:30 Corpus of usage examples: What is it good for? Timofey Arkhangelskiy

A Preliminary Plains Cree Speech Synthesizer. Atticus Harrigan, Antti Arppe and Timothy Mills

12:00–12:30 A biscriptual morphological transducer for Crimean Tatar. Francis M. Tyers, Jonathan Washington, Darya Kavitskaya, Memduh Gökırmak, Nick Howell and Remziye Berberova

12:30–14:00 ***Lunch***

Second afternoon: Language technologies – speech and morphology

Wednesday, February 27th, 2019 (continued)

14:00–14:30 **Improving Low-Resource Morphological Learning with Intermediate Forms from Finite State Transducers. Sarah Moeller, Ghazaleh Kazeminejad, Andrew Cowell and Mans Hulden**

14:30–15:00 *Bootstrapping a Neural Morphological Generator from Morphological Analyzer Output for Inuktitut*
Jeffrey Micher

15:00-15:30 **Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a Finite-State Transducer. Lane Schwartz, Emily Chen, Benjamin Hunt and Sylvia L.R. Schreiner**

15:30–16:00 ***Break***

16:00–17:00 ***Discussions, Looking ahead: CEL-4 and new ACL Special Interest Group***

Towards a General-Purpose Linguistic Annotation Backend

Graham Neubig[†], **Patrick Littell**[‡], **Chian-Yu Chen**[†], **Jean Lee**[†],
Zirui Li[†], **Yu-Hsiang Lin**[†], **Yuyan Zhang**[†]
[†]Language Technologies Institute, Carnegie Mellon University
[‡]National Research Council Canada
gneubig@cs.cmu.edu

1 Introduction

Language documentation is inherently a time-intensive process; transcription, glossing, and corpus management consume a significant portion of documentary linguists' work. Advances in natural language processing can help to accelerate this work, using the linguists' past decisions as training material, but questions remain about how to prioritize human involvement.

In this extended abstract, we describe the beginnings of a new project that will attempt to ease this language documentation process through the use of natural language processing (NLP) technology. It is based on (1) methods to adapt NLP tools to new languages, based on recent advances in massively multilingual neural networks, and (2) backend APIs and interfaces that allow linguists to upload their data (§2). We then describe our current progress on two fronts: automatic phoneme transcription, and glossing (§3). Finally, we briefly describe our future directions (§4).

2 Overall Framework

The final goal of our project is to create a linguistic annotation backend (LAB), that will take in raw or partially annotated linguistic data, and provide annotation candidates for a linguist (or other interested party) to peruse. Candidates for the types of services to provide are automatic phoneme transcription (Adams et al., 2018; Michaud et al., 2018), speech-to-text alignmen (Johnson et al., 2018), word segmentation (Peng et al., 2004; Goldwater et al., 2009), morphological analysis (Yarowsky and Wicentowski, 2000), syntactic analysis (Nivre, 2005), automatic glossing (Riding, 2008), or linguistic typology prediction (Daume III and Campbell, 2007). The LAB will be hosted on a server and exposed as an API that can be linked to popular annotation software such as ELAN[1] or FLEx.[2]

The obvious difficulty in creating such an interface is data scarcity in the languages in question. In order to overcome these barriers, we plan to take advantage of recent advances in NLP that allow for multilingual modeling (Täckström et al., 2012; Johnson et al., 2016) and multi-task learning (Caruana, 1997), which allow models to be trained with very little, or even no data in the target language (Neubig and Hu, 2018). We also plan to utilize active learning (Settles, 2009), which specifically asks the linguists to focus on particular examples to maximize the effect of linguists' limited time when working with field data. While there is still no alternative to significant human engagement when processing data, many of the decisions a linguist is faced with when transcribing, glossing, organizing, or searching a corpus are relatively rote – decisions that could be deducible from past decisions or from similar languages.

3 Current Progress: A Backend/Interface for Automatic Phoneme Transcription and Glossing

As first steps towards realizing our final goal, we have currently developed a backend for two tasks (automatic phoneme transcription and glossing), which is integrated with a simple example interface.

3.1 Backend Overview

The current LAB is based on a simple three-step process:

Data Upload The linguist uploads any existing annotated data to the interface.

[1] https://tla.mpi.nl/tools/tla-tools/elan/
[2] https://software.sil.org/fieldworks/

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 1–4,
Honolulu, Hawai'i, USA, February 26–27, 2019.

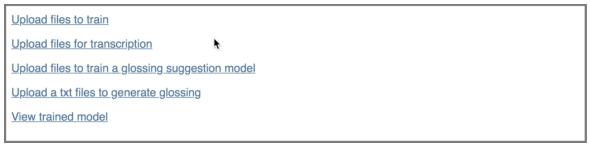

Figure 1: The prototype of the automated interface supporting transcription and glossing.

Figure 2: Uploading the training data to the automated interface for transcription model training.

Model Training A model is trained to process this data. This training could potentially utilize other data sources for multilingual and multi-task training.

Data Annotation The linguist uploads unannotated data, and the trained model proposes annotations for the linguist to accept or edit.

An example of an overall interface exposing this functionality for the currently implemented tasks of transcription and glossing is shown in Figure 1.

3.2 Phoneme Recognition

The automatic phoneme transcription component provides an interactive online interface for users to manage speech recognition models and transcribe speech recordings. The speech recognition model can be any one of the user's choosing as long as it supports the API. In our current system, we use Persephone (Adams et al., 2018) as our transcription backend, which is designed for low-resource language transcription. Through the API, the users can upload a batch of speech recordings along with the corresponding transcriptions as the training data to train a transcription model tailored to the language and speakers of their interest. The system is equipped with some default model and training configurations so that the users are not required to have expert knowledge of the transcription model and training. The model obtained from each training session will then be stored for later use. Figures 2 to 4 show the work flow of training a transcription model.

The users can upload speech recordings they want to transcribe to the interface, and perform the automatic transcription using previously trained models (Figures 5 and 6). The interface shows the transcription results to the users, and the users can

Figure 3: Training a transcription model using the training data uploaded by the user.

Figure 4: Training a transcription model using the training data uploaded by the user.

optionally edit the transcription results to fix errors or make model improvements (Figure 7). The refined transcriptions can then be downloaded by the users. If the user's data privacy preferences allow, the system can also collect them along with the original speech recordings as extra training data to further fine tune the model.

3.3 Automatic Glossing

The interface also supports making glossing suggestions. Glosses are generated word-by-word with Moses (Koehn et al., 2007), a statistical machine translation system. The system takes parallel data as input, which could be either the language and translations, or the language and glosses. Using this parallel data, we learn a word alignment with a statistical model, specifically the IBM alignment models (Brown et al., 1993) as implemented in GIZA++ (Och and Ney, 2003). Then we perform phrase extraction (Koehn, 2010), which gives us a translation probability distribution for each word or phrase in the combined corpus. We then display translations with high probability as glossing suggestions. An example of how the automatic glossing suggestion works on the interface can be seen in Fig. 8.

4 Future Plans

Working with field data is highly rewarding, but on a moment-to-moment basis the work is not usually particularly *engaging*; most of the individual decision events that a linguist makes during field corpus creation do not fully engage their reasoning capacity. Our goal is to maximize the effects of human engagement with data by maximizing the time the linguist spends on interesting and relevant

Figure 5: Uploading the speech recordings to transcribe.

Figure 6: Transcribing the speech recordings using the model previously trained.

decisions. We intend to explore this question with respect to both low-level decisions ("What word was said here?") and high-level decisions ("These utterances exemplify ergativity in this language; are there other examples in this corpus?"). Our future work towards this goal will take a three-pronged approach: developing a general-purpose linguistic annotation API and integrating it with popular annotation frameworks, developing new methods to perform multi-lingual and multi-task learning to train effective models even in a paucity of training data, and working with linguists to help refine and prioritize our work in these areas. In particular, for the third goal we are actively seeking collaborators who would be interested in testing and giving advice about the utility of the proposed approach.

Acknowledgements

We thank Alexis Michaud for his useful comments and help in preparation of data, Oliver Adams for his help with Persephone, and Antonis Anastasopoulos for helping us access and prepare the Griko data. This material is based upon work supported by the National Science Foundation under Grant No. 1761548.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Language Resources and Evaluation Conference (LREC)*., Miyazaki, Japan.

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*.

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–312.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Hal Daume III and Lyle Campbell. 2007. A Bayesian model for discovering typological implications. In *Proc. ACL*, pages 65–72. Association for Computational Linguistics.

Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1).

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation and Conservation*.

Melvin Johnson et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Philipp Koehn. 2010. *Statistical Machine Translation*. Cambridge Press.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proc. ACL*, pages 177–180.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Sverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with na data and the persephone toolkit. *Language Documentation and Conservation*.

Graham Neubig and Junjie Hu. 2018. Rapid adaptation of neural machine translation to new languages. In *Proc. EMNLP*, Brussels, Belgium.

Joakim Nivre. 2005. Dependency grammar and dependency parsing. *MSI report*, 5133(1959):1–32.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Fuchun Peng, Fangfang Feng, and Andrew McCallum. 2004. Chinese segmentation and new word detection using conditional random fields. In *Proc. COLING*.

Jon D Riding. 2008. Statistical glossing, language independent analysis in bible translation. *Translating and the Computer*, 30.

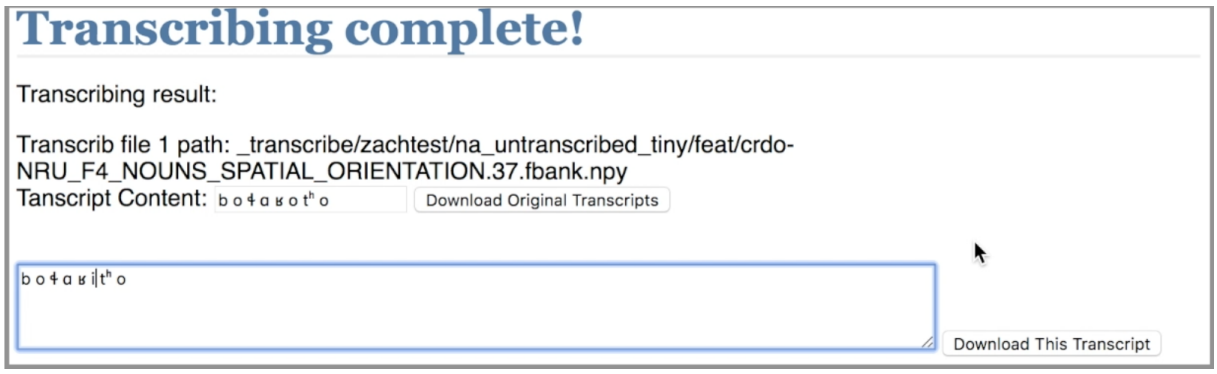

Figure 7: The users can examine the transcription results and optionally edit the results to correct errors.

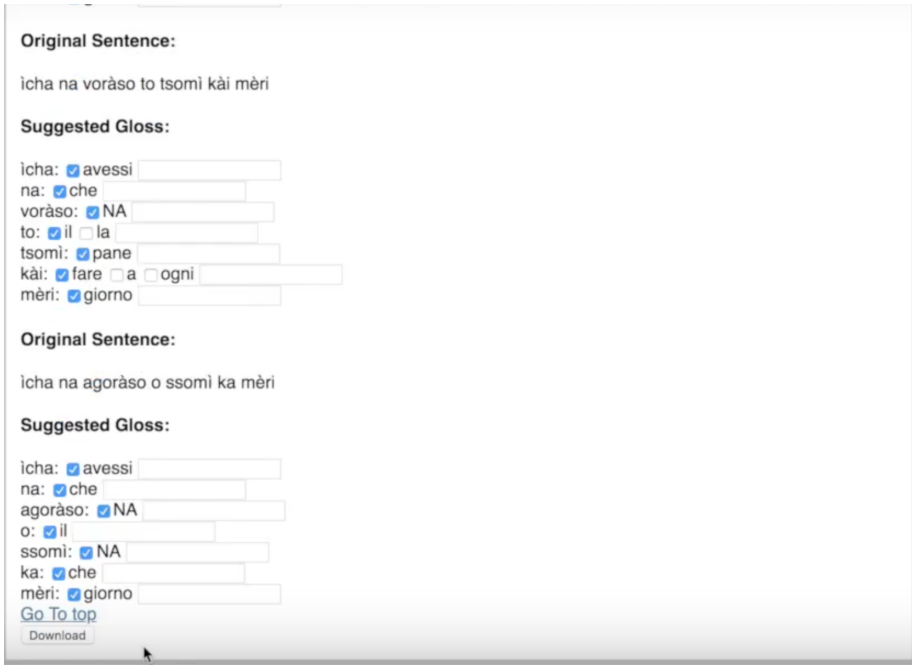

Figure 8: An example of generated glosses, for Griko language data from Anastasopoulos et al. (2018).

Burr Settles. 2009. Active learning literature survey. Computer Sciences 1648, University of Wisconsin–Madison.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proc. NAACL*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

David Yarowsky and Richard Wicentowski. 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proc. ACL*.

Building a Common Voice Corpus for Laiholh (Hakha Chin)

Kelly Berkson[†], **Samson Lotven**[†], **Peng Hlei Thang**[†], **Thomas Thawngza**[†], **Zai Sung**[†], **James C. Wamsley**[†], **Francis Tyers**[†], **Kenneth Van Bik**[‡], **Sandra Kübler**[†], **Donald Williamson**[†], **Matthew Anderson**[†]

[†] Indiana University, [‡] California State University Fullerton

[†] {kberkson,slotven,phthang,tzthang,zhsung,jwamsley,ftyers,skuebler,williads,andersmw}@indiana.edu,

[‡] kvanbik@exchange.fullerton.edu

Abstract

In this paper, we discuss our efforts to build a corpus for Laiholh, also called Hakha Chin. Laiholh is spoken in Chin State in Western Myanmar, in parts of India and Bangladesh, and in several Burmese refugee communities in the US. Indiana, for example, is home to about 25,000 Burmese refugees. The ultimate goal of our team is to contribute to the development of speech translation technology that will be of benefit, both in general and in the local community in Indianapolis. Translation tools would be of great use in local emergency rooms, schools, and businesses. In pursuing our (admittedly lofty) goals, we are building a growing community of speakers, field linguists, computational linguists, and computer scientists. As a team, we have worked to share our different skill sets and mobilize the wider community around collecting data via Mozilla's Common Voice platform. We present here a reflection on the project thus far, the kind of description we wish had existed when we were first building this collaboration and determining preliminary project goals. We hope that other communities and language activists who are thinking about developing speech technology may benefit from hearing about our motivations, concerns, experiences, and successes.

1 Introduction

One of the largest incoming refugee groups in the US is from Myanmar (Burma). Since 2008, more than 109,000 Burmese refugees have settled in the United States (Centers for Disease Control and Prevention, 2016), and at present admission rates for Burmese refugees are second only to the Democratic Republic of the Congo (DRC). From October 2017 through September 2018, 3,555 people from Myanmar and 7,878 from the DRC have been admitted to the US. (Bureau of Population and Migration, 2018).

Many Burmese refugees in the US are members of the Chin ethnic group, and the number of Chin churches, businesses, and community organizations is on the rise. Indiana is home to organizations such as the Burmese American Community Institute (BACI) and the Chin Community of Indiana (CCI), which focus on community support, integration, and advocacy. The BACI has a specific goal of preparing Burmese American students for higher education, e.g. with summer college prep and research programs, and their efforts have had growing success. Many graduates of their summer research program—including three of the authors of this paper—have enrolled at Indiana University, Bloomington as undergraduates. These circumstances have led to many opportunities for interaction and collaboration between linguists and students, a situation which is certainly echoed at many other schools around the world.

Members of the Chin communities in the US speak 30 or more under- and un-documented languages from the Kuki-Chin branch of the Tibeto-Burman language family. One of these languages is Laiholh, also sometimes called Hakha Lai, Lai Chin, or Hakha Chin (Bedell, 2001; Matisoff, 2003; Peterson, 2016; Van Bik, 2006). Laiholh is used as a vehicular language in Chin State and is spoken by as many as 10,000 people in Indiana—including four of the authors of this paper—as either a first or second language (Executive Director of the Burmese American Community Institute, 2018). Many other languages (e.g. Falam, Lautu/Lutuv, Mara, Matu, Zophei) are also spoken, albeit often by fewer people. Laiholh is not endangered, and we believe that will remain true despite the disruptive nature of many of the factors contributing to the formation of the diaspora communities in Indiana and beyond. Our hope, however, is that we can make a methodological contribution to communities working with smaller and endangered languages by sharing details about the

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 5–10,
Honolulu, Hawai'i, USA, February 26–27, 2019.

participatory and collaborative nature of our work. The work we describe here focuses on Laiholh, for purely strategic reasons: given the current composition of the community in Indianapolis, our hope is that Laiholh resources will have the greatest effect. We also hope to replicate our efforts with other Chin languages in Indiana, including endangered languages, in future work.

In this paper we describe one specific collaboration, but our central concern is not unique to us: we seek to build tools that will enhance communicative options. As noted by the United Nations High Commissioner for Refugees, the world refugee population is higher than it has ever been: someone is displaced every two seconds. Situations may develop quickly, and refugee communities around the world face communicative challenges. Existing literature focuses on communication challenges in, e.g., medical settings (Carroll et al., 2007; Morris et al., 2009) and primary and secondary schools (MacNevin; Naidoo, 2011). Language learning, even when proceeding well, takes time, and developing automatic solutions is even more time consuming. It is sometimes difficult to determine how we can best be of assistance in the face of pressing needs because the path to usable products is long. Herein, we outline one way in which student speakers, community members, linguists, and computer scientists can work together to begin to do so.

2 Community Needs

In developing our community of collaboration, one explicitly stated goal was to actively seek consensus on the projects we take on. We want to be of use, and the discovery and articulation of community needs has required reflection on the part of the native speakers and active listening on the part of the rest of the team. Discussions with team members, members of the wider Burmese community, and those who interact with them (e.g interpreters, speech pathologists) have allowed us to compile a list of the varied, real, and current needs of the local Burmese refugee community. Many of these needs revolve around language–a reality echoed in many other refugee communities worldwide.

Many challenges can be mediated with the help of a human interpreter, and the native speaker authors of this paper are often asked to interpret for family and community members. As such, they have firsthand knowledge of situations where translation is needed. Examples include:

- at the hospital and dentist (during check-in/check-out, before translators arrive)
- when paying bills (e.g., utilities) or interacting with insurance agents (e.g., car accidents)
- at state/government offices like the Bureau of Motor Vehicles (address changes, license plate renewal, ID card creation) or post office (for address changes, sending/reading mail)
- at car dealerships or local businesses (negotiation, sales)
- in interactions with the police (e.g. when pulled over), in court (pre-trial hearings), in jail (paying bail, calling a lawyer)
- at work (understanding/negotiating contracts/policies, talking to HR, requesting time off through FMLA, training)
- at school (parent-teacher interactions, administrative messages, meetings with speech pathologists)
- learning local regulations, e.g. Dept. of Natural Resources (hunting/fishing laws, licenses)
- with banks and credit card companies (understanding policies, paying bills)
- with the city government (e.g., to request building permits, pay parking fines)
- with US Citizenship and Immigration Services (citizenship paperwork, in-person interactions)
- for voting, voter registration, candidate info.

Professional interpreters can be employed to help in many situations—events such as court appearances require the human interpretation skills of a trained professional, for instance—but there are many situations where calling an interpreter is either not practical or not possible. Our student co-authors and others of their generation often interpret for community members, but they also attend school and work part-time so their availability is limited. If a community member cannot be accompanied by a bilingual interpreter to pursue a change of address at the Bureau of Motor Vehicles, a note saying "I need an address change form" may suffice. Such a solution fails as soon as a follow up question is posed, however, and in these mundane but crucial situations, technology could be of use.

There are also time-sensitive situations where technology could make a critical difference in the lives of the Burmese American community– in emergency rooms, for example, there is generally a lag between when patients check in and when interpreters arrive. In Indianapolis, experience suggests that often a Burmese translator (instead of a Laiholh translator) arrives, which is a problem because many Laiholh-speakers do not speak Burmese. Thus, while in some situations summoning the wrong translator might constitute only an annoyance, playing guessing games with interpreter language in an ER can be deadly. Simply put, the scale of the need (25,000 refugees in Indiana alone) requires what is currently an unreasonable amount of work for humans. As such, speech translation would be a boon—for many reasons, to many people, in our community and in others worldwide. It is against the backdrop of these realities that we, the authors, came together to form a collaborative team of speech community members, linguists, and computer scientists.

3 A Developing Collaboration

In December of 2017, a subset of the authors had a series of meetings to discuss collaboration concerning developing machine translation and automatic speech recognition (ASR) capabilities for under-resourced languages. We chose the development of a Common Voice system for Laiholh (see below) as the first step towards the larger aim. Since that time, our circle of collaborators has grown. The native speakers involved in this project went on to work as language assistants in a Field Methods class on Laiholh and on the Common Voice project described here. Several field methods students have continued to work on Laiholh. Input from computational linguists and computer scientists has informed the way field linguists collect, organize, and prepare data. Dialogue between all parties has been ongoing, and many of us work closely with one another on a near-daily basis.

Soon after our initial meetings, word of our interest in Chin languages spread and community members (both those in the Chin community and those who interact with them) began to contact us to describe challenges they encounter. Speech language pathologists from a local school district expressed a need both for basic materials on the languages spoken by their nearly 5,000 Burmese students and for help determining which language(s) children are acquiring at home. We talked to doctors who had little way to interact with Chin patients and difficulty providing them with written materials, and heard stories suggesting occasional patient discomfort when translators were involved in sensitive medical conversations. We met with the members of the Myanmar Students' Association on campus and learned that dozens of students were interested in working to develop language materials. As our list of needs and resources grew, it became clear that we needed a larger structure for data collection and a platform designed for inclusion, so that we could involve many eager parties. We needed to organize a group of individuals with different skill sets and different backgrounds around a common project, one with the potential to push us towards our growing list of goals.

We now provide a brief overview of Common Voice, noting why we believe its corpus-building structure and potential to lead to the development of voice recognition technology will help move us toward our larger goals.

4 Common Voice

Our long-term goal is to develop automatic speech translation for Laiholh, and Mozilla's Common Voice platform[1] offers two necessary outcomes that bring us closer to that goal. It facilitates both the creation of a public domain spoken corpus and the development of speech-to-text software. Speech data is collected via a phone or browser app from any native speaker willing to donate their voice to the corpus. Once the corpus is large enough, Mozilla will use machine learning software to develop speech-recognition technology This moves us closer to our larger goal of speech translation because once a written Laiholh sentence can be generated from spoken language, that written language can be used as input for text-based (Laiholh–English or English–Laiholh) machine translation technologies.

There are four specific ways that using Common Voice facilitated our work: (1) it provided us with a clear project structure and delineated, attainable goals; (2) it gave us an existing interface for data collection so we did not have to create one from scratch; (3) it stores the data collected, so we do not have to secure storage space for hundreds of hours of audio; and (4) it offers access to machine learning technology in the creation of a speech recognition system, which is a prerequisite for machine translation.

[1] https://voice.mozilla.org/

Before speech data collection can begin with Common Voice, two projects have to be completed. First, the Common Voice interface must be translated into the target language. For Laiholh, we completed this in Summer 2018. Next, 5,000 written sentences (in the target language) have to be collected. These written sentences are presented for users to read aloud, meaning that literacy is an important component of interacting with Common Voice. As such it may not be viable when working with languages that do not have widely-used orthographies.

Construction of a written corpus for Laiholh consisting of 5,200 sentences was completed in October 2018. The bulk of the work was completed by our native speaker undergraduate co-authors, who spent many hours a week in Summer 2018 thinking up sentences. As a larger group, we also sat together and thought through scenarios: "What do we need to say during parent-teacher conferences? What questions might a doctor ask at a check-up? What do we say when we're texting with friends?" Many sentences were also gleaned from an online Laiholh dictionary created by one of our co-authors, Kenneth Van Bik, and Laiholh author Joel Ling gave us permission to borrow sentences from some of his books. Finally, we also asked for input from other community members when translating some of the terminology in the interface to ensure that the decisions we were making on a day-to-day basis would result in an app that is user-friendly for everyone.

Common Voice offers a simple user interface where speech community members can provide two types of data: (1) Recording, where users are presented with a series of sentences which they are instructed to record; or (2) Validation, where users see a sentence, hear a recording from another user, and assess whether what they saw matches what they heard. Over time, as people submit and validate recordings, a large data set is collected and housed by Mozilla. The data set remains public domain and can be downloaded at any time for our own use, or by others. The data is structured as written sentences paired with multiple audio files and judgments of which audio files are accurate renderings of the provided sentences.

To ensure the development of robust technology, particularly given that Laiholh is spoken by a multilingual diaspora community, our goal is to record highly varied data that includes many accents, dialects, and voice qualities. This will ensure that we can train a robust speech recognition system. The only way to procure such a set of learning data is through widespread community use of the Common Voice app. This, in turn, means that we need to find ways to disseminate information about the project to as many people in the wider community, those beyond our smaller community of collaboration, as possible.

4.1 Preliminary Data

We began the speech data collection stage of the project in mid-November 2018. During the first three months of data collection—from November 14, 2018 to January 14, 2019—260 users have contributed by donating their voices and 310 users have contributed by validating sound clips. Altogether, 4,500 audio clips totalling 5 hours and 53 minutes of audio data have been submitted and 2 hours and 44 minutes have been validated. The average number of clips contributed per user is 17.4, and the top user contributed 243 clips.

To the best of our knowledge, this dataset constitutes the largest spoken corpus of Laiholh in existence. The Chin Cable Network Channel (CCN) made a video tutorial in Laiholh on how to use the app which was shared on CCN's Facebook page. We have also been invited to share brief presentations about the app during events at local churches.

While the total amount of data collected continues to grow day by day, growth of the corpus is clearly most robust when we are actively working to publicize it. During the first week that it was live, for example, more than 2 hours' worth of data were recorded. Community buy-in is hugely important in this work, and one request we received during the early weeks of data collection had to do with the style of the sentences included in the corpus. In particular, we were asked to add additional sentences that represented more informal domains such as texting and online chatting. In response to this request, we have pulled back from publicizing for the time being in order to devote time to increasing/diversifying the sentences in the corpus.

5 Developing Trust

In building a community based on different skill sets and different understandings, we have found two parts of the process where we have had to place trust outside of ourselves in order to pursue the project with Common Voice. First, we needed to trust that the Common Voice platform would function as advertised and that, if we dedicated time and resources to it, we would eventually get the desired output. Second, we needed

to trust that the wider Laiholh-speaking community would be able and willing to access Common Voice and record sentences. We turn now to describing why our developing community decided to place its trust in this project and why we are hopeful that we are on the right path.

We decided to pursue Common Voice because we were able to read online about other communities who were involved in the project. We read blog posts about various language groups hosting Common Voice "sprints" focused on collecting a lot of data in a short period of time, and we could envision doing that with our community. We read about the ethics and principles espoused by Common Voice.[2] We liked the emphasis placed on open access and public domain resources. One of our co-authors had inside knowledge of Common Voice, and talked with the team extensively—answering questions, and sharing information. Everything we heard was to our liking, and we felt comfortable moving forward.

With regards to the Laiholh speaking community and our questions about whether people would be interested in donating their time and voices to the project, we were able to put our concerns to rest very early on due to the enthusiasm we encountered. From college students to community leaders, we were met with excitement and interest everywhere we went. Organizations like the Chin Cable Network Channel and the Chin Youth Network of North America have helped us by creating video tutorials and advertising videos. A representative of Chin Baptist Churches USA (CBC-USA) offered to advertise the project to all of its 110 churches across the country with its 30,000 or so members. These positive reactions, in addition to the response when Common Voice in Laiholh went live, reassured us that with continued effort on our part we will continue to see robust community participation in data collection.

6 Conclusion

Seeking to develop speech recognition and machine translation technology is a sizeable goal that involves many steps. To accomplish this goal, we will need to use all of our diverse strengths and skills, and we will need to develop a common "language" that will allow researchers from computational linguistics, computer science, linguistics, and from the language community to develop our capacity to collaborate and to trust in one another. One challenge that we continue to work to confront has to do with communicating complex facts to non-experts: the language experts on our team have knowledge about Laiholh that can be difficult to convey but crucial for the computer scientists to understand. Similarly, the long and complicated path that we hope will result in working speech technology for Laiholh—and the way in which specific components of the project like building the Common Voice corpus are related to that larger goal—is clearer to the computer scientists and computational linguists than to the other members of our team, or indeed to other members of the larger community. To maintain energy and enthusiasm, however, it is crucial for the technical experts to find ways to make the steps more transparent. Working to ensure that all members of the team are engaged and empowered is an ongoing goal, one that has been well-served by coming together around the shared Common Voice project.

Our goals are lofty, but the payoff if we succeed will also be very high as it will dramatically improve the lives of local, national, and international community members. The hope, too, is that successful collaboration will increase our knowledge about how speech recognition and machine translation can work for other languages with few computational resources but with a strong community buy-in.

[2]For example, see the following blog post: https://medium.com/mozilla-open-innovation/more-common-voices-24a80c879944

Acknowledgments

This project was supported in part by the Indiana University College of Arts and Sciences Ostrom Grants Program and by the IU Department of Linguistics.

References

George Bedell. 2001. The syntax of deixis in Lai. *Linguistics of the Tibeto-Burman Area*, 24(2):157–171.

Elaisa Vahnie. Executive Director of the Burmese American Community Institute. 2018. Private Communication.

Jennifer Carroll, Ronald Epstein, Kevin Fiscella, Teresa Gipson, Ellen Volpe, and Pascal Jean-Pierre. 2007. Caring for Somali women: Implications for clinician–patient communication. *Patient Education and Counseling*, 66(3):337–345.

Centers for Disease Control and Prevention. 2016. Burmese refugee health profile. Technical report, U.S. Department of Health and Human Services.

Joanne MacNevin. Feeling our way in the dark: Educational directions for students from refugee backgrounds. Master's thesis, University of Prince Edward Island.

James A Matisoff. 2003. *Handbook of Proto-Tibeto-Burman: System and Philosophy of Sino-Tibetan Reconstruction*. University of California Press.

Meghan D Morris, Steve T Popper, Timothy C Rodwell, Stephanie K Brodine, and Kimberly C Brouwer. 2009. Healthcare barriers of refugees post-resettlement. *Journal of Community Health*, 34(6):529.

Loshini Naidoo. 2011. What works? A program of best practice for supporting the literacy needs of refugee high school students. *Literacy Learning: The Middle Years*, 19(1):29–38.

David A Peterson. 2016. Hakha Lai. *The Sino-Tibetan Languages*, page 258.

Refugees Bureau of Population and Migration. 2018. Refugee arrivals by placement state and nationality. Technical report, U.S. Department of State.

United Nations High Commissioner for Refugees. 2018. Refugee statistics. Technical report, United Nations Refugee Agency.

Kenneth Van Bik. 2006. *Proto-Kuki-Chin*. Ph.D. thesis, University of California, Berkeley.

Digital Dictionary Development for Torwali, a less-studied language: Process and Challenges

Inam Ullah
Torwali Research Forum
Bahrain, Swat, Khyber Pakhtunkhwa
Pakistan
torwalpk@yahoo.com

Abstract

Torwali is an endangered and less-studied language spoken in the north of Pakistan. Recently, the community celebrated publication of the first ever Torwali dictionary both in print and an online version. This paper discusses issues and challenges regarding lexicography of a previously non-written language; from data collection by the native speakers having no set goals and training or institutional support, to organization and presentation of the data for producing multiple versions of the dictionary. The first section describes the process of developing the database using the methods of wordlists and semantic domains. The proceeding sections describe the technical development of its printed and online version in detail, and discuss orthographical, technical, computational and social concerns of the project. The paper concludes with recommendations for future dimensions of the present work and for similar projects with special consideration to lexicographical work on non-written languages.

1. Introduction

1.1 Torwali language

Torwali belongs to the Kohistani sub-group of the Indo-Aryan Dardic languages, spoken in the upper reaches of district Swat of northern Pakistan. It has two dialects (the Bahrain and Chail dialects), with a total of approximately 90,000 to 100,000 speakers. Close to half of the population has migrated to bigger cities where language shift is a common phenomenon.

1.2 Motivation or need for the project

The project initiator, a mother-tongue speaker of Torwali, when studied the written materials on the language for the first time, found many semantic and phonetic errors. This initially motivated him to work on his native language in order to present it more accurately to the academic community. Later, after receiving encouragement from the community elders, he decided to compile a dictionary of Torwali based on the idea that dictionaries can be a crucial resource for language learning and instruction, particularly with regard to endangered languages. A good dictionary can address issues of orthography, documentation and language preservation. Previously, the locals found it difficult to write Torwali language using the alphabets of neighbouring regional languages or the national language as some of its peculiar sounds had no representation in their alphabets. The main goals were, therefore, to record, document and preserve a hitherto unwritten language of Swat Kohistan and thus, to safeguard it for the future generations.

1.3 Intended audience of the dictionary

Initially, the intended audience was the academic community. The aim was to provide them with error-free material of Torwali for further research. However, later, in view of the interest of the Torwali community, it was decided that the intended audience would include both the academic and the speech communities. During the compilation process, numerous difficulties

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 11–16,
Honolulu, Hawai'i, USA, February 26–27, 2019.

emerged regarding decisions to present the data in a way that would benefit both the communities equally. As a result, it was decided that the primary audience would include those Torwali-speaking Torwalis and Torwali-learning Torwalis whose preference is the socio-cultural information like clans, place names, medicinal plants, cultural items, myths and oral traditions. Thus, the final product of the database is intended for students, Torwali speakers across the globe, tourists, and researchers.

1.4 The selection of dialect

While compiling data for Torwali dictionary the 'Bahrain dialect' was decided to be the standard dialect because: (i) it is spoken by a larger number of Torwali speakers; (ii) Bahrain is the cultural, political and administrative center of Torwali community; and, (iii) the compiler of the data speaks Bahrain dialect of Torwali.

Despite the above-mentioned decision, some words peculiar to the Chail dialect were added to the database with the tag of 'Chail dialect' However, it was not possible to enter Chail variation of every Torwali word due to space issues.

1.5 Previous literature

Several western researchers have worked on this language. In 1880, John Biddulph published *Tribes of the Hindoo Koosh*, which contained the first linguistic description of the Torwali language. The most extensive work on the language was carried out by Sir George Grierson which is known as *Torwali: A Dardic language of Swat Kohistan* (1929). In the late 1980s, SIL International carried out a sociolinguistic survey in northern Pakistan which included the Torwali community. Wayne Lunsford's work, *An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan* (2001) is another major work on Torwali after Grierson's.

2 The undecided project goals

There were no set goals at the beginning of the project. It was all about 'writing a dictionary of Torwali'. The compiler, being a government school teacher, had no previous knowledge or training of lexicography. He worked on the project as a hobbyist and therefore, did not time-frame it. However, he stored the database in an electronic format to serve the academic purposes of research.

3 Methodology: Printed Version

Major part of the database was developed over the past one and half decade by the active help of the author's students, colleagues, friends and relatives. Data was collected on index cards and paper and was entered in to the computer program called 'Shoebox'. 'Shoebox' was replaced with the improved version 'Toolbox'

3.1 Data collection and verification

Both wordlist and semantic domain methods were used for the data collection. Being bilingual, the author used Urdu wordlists for recalling Torwali words. But specific cultural items, plants and animal names about which the author himself was unaware could not be recorded using this method. He, therefore, adopted the method of semantic domains. He asked his Torwali students, friends and family members to make lists of words relating to a specific semantic domain or sub-domain. For example, a group of students was asked to bring a labeled sketch of the interior of a watermill.

To ensure correctness and completeness, cross-checking and verification of the data was conducted through multiple sources within the community, such as, various people living in different localities (valleys and side-valleys of the indigenous area as well as in different urban centers).

3.2 Expansion of the database

Printouts of the existing database were distributed among the Torwali speakers for verification and further addition of lexical items that had been left out. At the same time the author made a partial use of lists of *Semantic Domains* prepared by SIL International under The Dictionary Development Process (DDP) http://www-01.sil.org/computing/ddp/ddp_wordcoll.htm?. This process facilitates lexicographers to collect words for the development of dictionaries of minority languages. It helped increase the database from 5200 to 8000 lexical entries by including names and related information of places, plants and clans as well as idioms, proverbs and, words related to Chail dialect.

3.3 Verification and refinement of specific semantic domains

Lists containing words of special domains were verified by the Torwali practitioners of the field concerned. For example, Torwali words for diseases and ailments were verified by and discussed with qualified medical practitioners living in the indigenous area. Similarly, items relating to forests, botany, watermills or agriculture were verified by the Torwali speakers working in the respective fields.

3.4 Consistency Checks

Consistency checks were employed both automatically and manually, depending on the availability of the features in the Toolbox program in which the database was stored and handled. For example, Toolbox consistency checks support parts of speech but not spell checks, particularly in the national and source languages. Thus, spell checks were carried out manually.

3.5 The use of lexique pro

Toolbox file opened in Lexique Pro and was exported to Microsoft Word in a standard dictionary format. Thus final export was made through Lexique Pro instead of Toolbox due to repeated problems faced while exporting. But there were lots of formatting issues which had to be fixed manually. For example, various Toolbox field markers, such as, \ue, \vr, \lt, \va and \ps were changed into Urdu script to meet the practical needs of developing Torwali-Urdu print version as they were not supported by Lexique Pro during export process. Proofreading of the entire exported file was done page by page. Arising technical issues were resolved through the 'trial and error' principle.

4 Methodology: Online Torwali Dictionary (OTD)

4.1 Overall architecture (OTD)

To develop the online version of the dictionary, the lexicon data was contained in a Toolbox database file. The Toolbox database file was taken to Lexique Pro to be converted into an xml formatted file. The corresponding xml file was exported, using MDF (Multi- Dictionary Formater). The xml file was used as input to the dictionary's website application which transformed the word detail into html formatted content for users.

4.2 Microstructure and macrostructure of the OTD

Lexique Pro generated xml file of lexicon data in LIFT (Lexicon Interchange FormaT) language. The detailed xml encoding format of LIFT formed the microstructure. The method of accessing information by the user formed the macrostructure.

Macrostructure and microstructure were navigable to some extent. Words were linked on the basis of parts of speech, and synonyms were cross referenced as navigable links deeming OTD to be termed as hyperlexica (Gibbon, 1999).To ensure that user accessed the dictionary information with ease, the following navigational ways were made available on the website:

- Torwali alphabet was enlisted on the website. Through this, user could have the list of words starting with the selected Torwali letter.
- User could navigate the words by syntactic categories, affixes or phrases. These indices would display a corresponding wordlist making it easier to find the word. Native users could ponder upon the list and find out if some words were missing. It could help teachers make a lesson plan to teach a certain category to Torwali-learners.
- Users could search any Torwali word by entering it using onscreen Torwali keyboard provided on the website. If a word existed in the lexicon, then the result would either be a single word or more than one word if its homonyms existed.
- Through any of the above three ways a user could reach a word or word list. The word detail would be displayed by clicking on the desired word. It would contain all the information fields already displayed in the hard copy of the dictionary except the usage field. ('Usage' depicts the geographical usage, obsolescence and vulgarity of the word.)
- Reverse lookup; Urdu to Torwali navigation for words is also one of the

features of the online Torwali dictionary. Urdu word list would help user traverse back to Torwali equivalent.

4.3 Orthography Issues

Since Torwali was an oral language until the start of the lexical compilation, the orthography issues constituted a major source of problem during the process of compilation.

4.3.1 Decision on the script

In order to be in cultural and historical harmony with the regional language Pashto and national language Urdu, it was decided with the help of the community activists that Perso-Arabic script would be used for writing Torwali, because both the Pashto and Urdu languages are written in Perso-Arabic script.

4.3.2 Characters for peculiar sounds

There are five distinct sounds in Torwali which are absent in both Pashto and Urdu. Grierson identified and mentioned these sounds in his work (Grierson, 1929). A Dutch phonetician (Baart, 1999) worked on the neighboring Kalami (or Gawri) language in the 1990's and presented them to the community people for approval. After a detailed discussion all of them recognized the five sounds (figure 3).

4.3.3 Standard spellings

During serial workshops organized for discussion on language issues among Torwali language activists, it was noticed that many words were being written with different spellings. This issue was also evident in the database where non-unique words were quite often found for the same lexical item. The participants decided to adopt the spellings which occurred with high frequency.

4.4 Technical Issues

4.4.1 Conversion of legacy fonts into Unicode fonts

After using 'Shoebox' program for many years, Torwali lexical data had to be shifted to its newer version 'Toolbox'. Hence, all the legacy fonts needed to be converted to Unicode supported fonts. Almost all the characters representing Torwali sounds were assigned Unicode positions except the Voiced Retroflex Affricate.

4.4.2 Torwali support in Nafees Pakistani Web Naskh

Center for Language Engineering, CLE (formerly CRULP) developed the Burushaski-Torwali-Khowar (BTK) font which is a character-based Nafees Pakistani Web Naskh Open Type Font in 2009. It was an extension of Nafees Web Naskh supporting several regional languages including Torwali in addition to Urdu.

4.4.3 Torwali keyboard development

In order to support Torwali characters to be typed easily along with Urdu characters, Torwali keyboard was developed by Center for Language Engineering, CLE (formerly CRULP). This keyboard was based on and similar to the Urdu Phonetic Keyboard so that the additional characters for Torwali Language could be typed easily along with the regular Urdu characters.

4.5 Issues relating to XML file

4.5.1 Some of the Torwali examples were not exported to xml format by Lexique Pro

According to LIFT, xv field in word entry contained examples in vernacular language and xe, xn, xr fields contained examples in English, national and regional languages. If xv did not exist, other example fields could not be exported to xml format. In case of Torwali dictionary, xv field contained examples in the form of IPA symbols and xr contained examples in Torwali language. In some cases xv field did not exist or in other words pronunciation of example sentences did not occur. Therefore dummy xv was inserted where xv was empty by using Toolbox, so that xml element corresponding to xr could be generated and thus displayed to the user. As xv-value was not to be displayed in hard copy or on website therefore dummy value could be used to save time and insert remaining pronunciations of Torwali example sentences afterwards.

4.5.2 Text formatting for hard copy dictionary

Toolbox was used to compile and manipulate the lexicon. However, its export features did not work well for publishing hard copy of the dictionary. For

this purpose Lexique Pro was used. There were default formatting styles (known as Multiple entry style) for each of the fields in a word entry. These styles were used by Lexique Pro during the process of export to HTML or WORD format.

4.5.3 Sorting of non-written languages

As Torwali was not a written language therefore collation sequence was not readily available for it. Though, collation rules had been explicitly mentioned in Toolbox, diacritics were not handled as ignorable characters due to which sorting was interrupted. The presence of diacritics caused the word to be processed in sub-sequences. Therefore, the hard copy of the dictionary was not properly sorted. This discrepancy was later removed and headwords were displayed in a proper sequence in the online dictionary version.

4.5.4 Gloss field and reverse lookup

In gloss field, semicolon is used to separate the multiple gloss terms. In Urdu gloss \gn, Urdu semicolon was used but was not recognized as a separator. Therefore, all the gloss field content was handled as single gloss term.

4.5.5 Encoding

Word detail of the Torwali-Urdu was in XML format therefore Unicode (that is, utf-8) encoding had to be used by the website. Secondly, Urdu and Torwali characters could not be presented by ASCII encoding. This is because web application configuration is set for languages using Unicode, otherwise the characters appear illegible on the interface.

4.6 Social and Other Issues

Like every living language Torwali has also many taboo and slang words. Torwali natives differed in their treatment of these words in the dictionary of their language. Some suggested that these words must be avoided as they may create wrong impression among the children and 'outsiders' about the community. Others said that these words were a part of their language and had to be recorded. After long sessions of discussions with community activists and elders it was decided that obscene slangs were to be avoided but words with offending connotations could be tagged with 'offending'. Similarly, some clans with shady histories did not want their historical information to become a part of the dictionary. Therefore, their names were included but not their history.

5 Future Work

Based on the existing database several enhancements can be made to enrich the practical uses of the dictionary.

Indexing on the basis of semantic domains can be incorporated in the interface.

The Torwali grammatical and collocation information can be enhanced to form a useful resource for Torwali to Urdu translational work leading to localization. When translating a sentence from source to target language, the context of the word sometimes changes the choice of word in target language (Saleem, 2007). Such constructions can be resolved when proper collocations and grammar of words are given.

The interactive interface for users can be added to contribute linguistic information of a new word or to an existing word. After the linguistic verification of the contributed information, it can either be approved and added to dictionary or disapproved.

Torwali to English dictionary can be developed and the corresponding online interface can be merged to the existing one - OTD.

Talking Torwali Dictionary can be produced to help community members living away from the indigenous area to learn the language of their ancestors.

Example sentences are good resource for better explanation of a word in a dictionary. There are only 1200 example sentences which need to be expanded under each lexical entry.

Specific information needs to be added for plants and animals rather than the generic formation as "a kind of…".

6 Recommendations

- The 'Usage' field should be exported by Lexique Pro. Currently, LIFT stated that <usage> element corresponding to \et exists but it is not exported by Lexique Pro.
- Collation sequence of non-written and less taught languages should be included by collation consortia, so that these can be readily available to linguistic tools. Collation sequence leads to properly sorted language data which is more efficient to navigate and manipulate.
- For non-written or less taught languages, data collection is quite a difficult task. Therefore, instead of top-down approach, bottom-up approach (Carr, 1997) is more helpful. Especially through emails or forums or a dictionary website page dedicated to user contribution are the easy and fast ways to collect the data. These small contributions can be of great benefit to all.
- Many cultural items, whose precise alternatives are not available in target languages, are best explained with the help of drawings and pictures.
- Idioms and proverbs embody the essence of a language. There are about 600 idioms and proverbs in the database which can fairly be expanded to thousands.

7 Acknowledgments

We want to thank all those who contributed in making possible the creation of Torwali dictionary: especially, the Torwali community who very persistently supported the compiler of this work, both practically and morally; National Geographic (http://www.nationalgeographic.com) who supported the completion and editing of the Torwali lexicon; the International Development Research Center (IDRC), Ottawa, Canada who funded the project for bringing the Torwali lexical data online; and the University of Chicago who initially supported the development of the dictionary content.

8 References

Baart, Joan L.G. 1999b. *'A Sketch of Kalam Kohistani Grammar'. Islamabad: National Institute of Pakistan Studies and Summer Institute of Linguistics. (Studies in Languages of Northern Pakistan Vol. 5).*

Carr, M. 1997. *'Internet Dictionaries and Lexicography.' Internal Journal of Lexicography, vol. 10 No. 1. 1 Feb 2011. http://ijl.oxfordjournals.org/content/10/3/209.full.pdf+html.*

Gibbon, D. 2000. *'Computational Lexicography.' In F. van Eynde and D. Gibbon (eds.), Lexicon Development for Speech and Language Processing. Dordrecht: ELSNET, Kluwer Academic Publishers, 1-42.*

Grierson, George. A. 1929. *Torwali: An Account of a Dardic Language of the Swat Kohistan. London: Royal Asiatic Society.*

Lunsford, Wayne. 2001. *An Overview of Linguistic Structures in Torwali, A Language of Northern Pakistan. M.A. Thesis, University of Texas at Arlington.*

Martin, J. B. & Mauldin, *M. M. 1997. 'Practical and Ethical Issues in Lexicography: Examples From The Creek Dictionary Project.' In C.Pye (eds.), 1996 Mid-America Linguistics Conference Papers, 565-573.*

Saleem, M. I. 2007. *'Bilingual Lexicography: Some Issues with Modern English Urdu Lexicography – a User's Perspective.' Linguistik online. 1 Feb 2011. http://www.linguistik-online.com/31_07/saleem.pdf/.*

Ullah, Inam. 2004. *'Lexical Database of the Torwali Dictionary.' In The Asia lexicography conference. Chiangmai: Payap University.*

Applying Support Vector Machines to POS tagging of the Ainu Language

Karol Nowakowski[*], Michal Ptaszynski[*], Fumito Masui[*], Yoshio Momouchi[**]

[*] Kitami Institute of Technology, 165 Koen-cho, Kitami, Hokkaido 090-8507, Japan
karol_nowakowski@ialab.cs.kitami-it.ac.jp, ptaszynski@cs.kitami-it.ac.jp, f-masui@mail.kitami-it.ac.jp
[**] Professor Emeritus of Hokkai-Gakuen University, Minami 26, Nishi 11, Sapporo 064-0926, Japan

Abstract

We describe our attempt to apply a state-of-the-art sequential tagger – SVMTool – in the task of automatic part-of-speech annotation of the Ainu language, a critically endangered language isolate spoken by the native inhabitants of northern Japan. Our experiments indicated that it performs better than the custom system proposed in previous research (POST-AL), especially when applied to out-of-domain data. The biggest advantage of the model trained using SVMTool over the POST-AL tagger is its ability to guess part-of-speech tags for OoV words, with the accuracy of up to 63%.

1 Introduction

Ainu[1] is a critically endangered language isolate spoken by the native inhabitants of northern parts of Japan. Due to its unique characteristics (such as noun incorporation or the usage of affixes – rather than pronouns – to express grammatical person), it has been the subject of a number of linguistic studies. Nevertheless, it receives little attention in the fields of NLP and Computational Linguistics. There is an ongoing project, started by Nowakowski et al. (2018), to create a large-scale annotated corpus of Ainu, which is expected to trigger further development of language technologies related to Ainu. However, there are few Ainu language experts, which renders the task of manual annotation very time-consuming if not infeasible. A possible solution to the problem is to apply bootstrapping techniques (as described e.g. by Clark et al. (2003)) in order to generate the annotations automatically or semi-automatically. As a starting point for such endeavor, in this paper we describe an experiment comparing the performance of two different automatic POS taggers on Ainu language data.

The remainder of this paper is organized as follows. In Section 2 we shortly describe the characteristics of the Ainu language. In Section 3 we review the related work. In Section 4 we introduce the data used to train the part-of-speech taggers applied in this research. In Section 5, the test data used in evaluation experiments is presented. In Section 6 we explain the modifications to part-of-speech annotations present in the data applied in our experiments and introduce the full POS tagset with statistics. In Section 7 we describe the SVMTool settings used for model generation and tagging process. Section 8 is dedicated to the evaluation experiments and discussion about their results. Finally, Section 9 contains conclusions and ideas for future improvements.

2 Characteristics of the Ainu language

In terms of typology, Ainu is an agglutinative language, with a tendency towards polysynthesis manifested by the presence of such traits as pronominal marking and noun incorporation (especially in the language of classical Ainu literature (Shibatani 1990)). The basic word order is SOV. Ainu verbs – and to lesser extent nouns –

kotan	apapa	ta	a=eponciseanu
kotan	apa-pa	ta	a-e-pon-cise-anu
village	entrance-mouth	at	we/people-for[someone]-small-house-lay

We built a small hut for [her] at the entrance to the village.

Figure 1: Example of polysynthesis in the Ainu language (Tamura 1996).

[1] The word *ainu* (written as *aynu* in modern standard transcription) means "human" and it is also used to refer to the ethnic group in question.

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 17–23,
Honolulu, Hawai'i, USA, February 26–27, 2019.

take a variety of affixes, expressing reciprocity, causativity, plurality and other categories.

History of Ainu as a written language is relatively short. Most documents are transcribed using Latin alphabet and/or Japanese *katakana* script (all textual data used in this research is written in Latin script). Until the last decade of the 20th century there existed no widely accepted standard orthographic rules for the Ainu language[2].

3 Related work

The first and hitherto the only existing part-of-speech tagging tool for the Ainu language was developed by Ptaszynski and Momouchi (2012), under the name POST-AL. It was trained using a dictionary of Ainu compiled by Kirikae (2003) and performed POS disambiguation based on word n-grams obtained from sample sentences included in the dictionary. In 2017, Nowakowski, Ptaszynski and Masui investigated the possibility of improving the system's performance by using two dictionaries instead of one and applying a hybrid method of part-of-speech disambiguation, based on word n-grams and Term Frequency.

Unlike POST-AL, state-of-the-art POS taggers developed for other languages typically utilize part-of-speech annotated language corpora as their training data. One of such tools is the SVMTool by Giménez and Márquez (2004), which is an open source generator of sequential taggers based on Support Vector Machines. It achieves an accuracy of 97.2% in POS tagging of English, but has also been applied in studies dedicated to low-resource languages, such as the ones by Hagemeijer et al. (2014) and Behera et al. (2015).

In this research we carried out an experiment to compare POST-AL and SVMTool. Specifically, we used SVMTool v. 1.3.2 (Perl version)[3] and POST-AL tagger in the variant with hybrid approach to POS disambiguation, which yielded the best results in experiments carried out by Nowakowski, Ptaszynski and Masui (2017).

There are several lexicons of the Ainu language containing information about parts of speech, such as those by Nakagawa (1995), Tamura (1996) and Kirikae (2003). However, the amount of existing POS annotated texts which could be readily applied as a training corpus for a tagging system is negligible. Nowakowski et al. (2018) have included three POS tagged datasets (less than 30 thousand tokens in total) in their corpus. In this research we use one of them – an online dictionary by Bugaeva and Endō (2010) – to produce training data for SVMTool (for details, see the next section).

4 Training data

To train both taggers used in this research, we used the data extracted from A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational Dictionary by Bugaeva and Endō (2010), which is an online dictionary based on the *Ainugo kaiwa jiten*, a dictionary compiled by Shōzaburō Kanazawa and Kotora Jinbō, and published in 1898. It contains 3,847 entries.

Apart from isolated headwords, the resource includes 2,459 multi-word items (phrases and sentences) and each of them is annotated with a sequence of POS tags. Using that information, we were able to build a small (12,952 token-tag pairs, excluding punctuation) part-of-speech annotated corpus. A subset of it was excluded from the training data, in order to be used as test data in evaluation experiments (for details, see the next section), which left us with a training corpus of 11,249 token-tag pairs (excluding punctuation).

In order to avoid an increase of Out of Vocabulary words, we decided to retain single-word entries in the training corpus and treated them as separate sentences (by inserting a sentence delimiter after each of them).

The corpus was prepared in column format (one token per line), which is the format accepted by SVMTool. Additionally, for the purpose of applying it with POST-AL, it was converted into a dictionary format, where each entry consists of a token (word or punctuation mark), part-of-speech and a list of sentences the given word appears in (if available). The resulting dictionary contains a total of 2392 entries.

5 Test data

To evaluate the performance of both taggers, we used two sets of held-out data:

TDOA: This dataset consists of 1701 tokens (excluding punctuation) from the A Talking

[2] Standard orthography has been proposed by the Hokkaidō Utari Kyōkai (1994) and is widely used to this day.

[3] The software and its documentation can be downloaded from http://www.lsi.upc.es/~nlp/SVMTool/

Dictionary of Ainu… (Bugaeva and Endō 2010). Samples for the test data were selected in the following way: firstly, all sentences with the token count (excluding punctuation) of 3 and higher were extracted from the training corpus and grouped according to their token count. Secondly, duplicate sentences were eliminated. In the next step, a random sample of 20% was selected from each group. Lastly, the sentences selected for the test data were excluded from the training corpus.

SYOS: Five out of thirteen *yukar* epics included in the *Ainu Shin'yōshū* ("Collection of Ainu songs of gods") by Yukie Chiri (1923). Unlike the A Talking Dictionary of Ainu…, it represents the literary style of Ainu. The text was revised in terms of transcription by an Ainu language expert. It comprises a total of 1606 tokens (excluding punctuation) in 88 sentences.

6 POS annotations and tagset

Before applying the annotations produced by Bugaeva and Endō in our research, we decided to introduce several modifications. All such decisions were consulted with three comprehensive dictionaries including the information about parts of speech, by Nakagawa (1995), Tamura (1996) and Kirikae (2003). We also referred to the classification of word classes proposed by Refsing (1986).

The most notable change is the elimination of two word classes: Numeral (135 occurrences in the original data) and Interrogative (347 occurrences). All tags belonging to these two classes were converted to one of the following tags, depending on morphosyntactic characteristics of words they denote: "Adnoun" (e.g. *sine* – "one [day]") or "Noun" (e.g. *sinep* – "one thing") for Numerals, and "Pronoun" (e.g. *hemanta* – "what"), "Adnoun" (e.g. *inan* – "which"), "Noun" (*hempakniw* – "how many people"), "Adverb" (e.g. *hempara* – "when") or "Locative noun" (e.g. *hunak* – "where") for Interrogatives. The reason for that modification is that, apart from Bugaeva and Endō only Nakagawa classifies such words simply as Numerals and Interrogatives, whereas both Tamura and Kirikae rely on functional criteria in deciding their primary word class. Apart from that, we corrected a number inconsistent annotations and typos, and annotated words for which POS tags were missing in the original data. Moreover, we added three punctuation marks that were absent from the A Talking Dictionary of Ainu…, but often appear in other texts: quotation mark (""), colon (":") and ellipsis ("…").

Tag	Number of occurrences	
	A Talking Dictionary of Ainu… (with modifications)	SYOS
Noun	2799	355
Intransitive verb	2504	297
Transitive verb	1503	174
Personal affix	1114	178
Adverb	1041	65
Conjunctional particle	626	146
Nominalizer	594	36
Locative noun	480	64
Final particle	430	22
Case particle	415	55
Adnoun	343	38
Postpositive adverb	246	8
Verb auxiliary	229	50
Supplementary particle	182	28
Pronoun	166	8
Ditransitive verb	130	18
Complete verb	56	2
Interjection	47	11
Proper noun	47	17
Prefix	0	3
.	3396	55
;	508	0
?	470	12
,	106	102
!	28	14
"	1	50
:	1	1
...	1	2
:--	0	5
Unknown	0	31

Table 1: Complete tagset and statistics.

Gold standard part-of-speech annotation for the SYOS dataset was performed by an Ainu language expert, in accordance with the methodology described by Momouchi et al. (2008).

The complete part-of-speech tagset along with statistics of occurrences in both datasets is presented in Table 1.

7 SVMTool settings

7.1 Model settings

Our model was trained on the column-formatted corpus described in Section 4, with training

parameters set to default values. Appendix A explains the feature set used in each variant of the model applied in this research.

Preliminary experiments revealed that the model assigns tags corresponding to punctuation marks (e.g. ":") to many lexical OoV words. To avoid such behavior, we modified one of the model files containing the list of tags to be considered for OoV tokens, removing such tags from the list.

7.2 Tagging parameters

In the experiments with SVMTool tagger, we investigated the performance with different values of the following parameters[4]:

- Tagging strategy (`-T`) – different strategies apply different tagging schemes (greedy or sentence-level) and different variants of the tagging model are used;

- Tagging direction (`-S`) – LR (left-to-right), RL (right-to-left) or LRL (both directions combined). According to Giménez and Márquez (2012), tagging direction "varies results yielding a significant improvement when both are combined".

8 Results and discussion

Results of POS tagging experiments using SVMTool for each combination of tagging parameters are shown in Tables 3 and 4, while Table 5 presents the results of experiments with POST-AL. Table 2 shows the MFT baselines calculated by SVMTool.

The results indicate that both taggers are better than the baseline and a tagger generated using SVMTool performs better than POST-AL, especially when applied to out-of-domain data (SYOS). The biggest advantage of the model trained using SVMTool is its ability to predict part-of-speech tags for Out of Vocabulary words, which it performs with the accuracy of up to 63% (see Tables 6 and 7), while POST-AL does not have such a mechanism. In fact, if we excluded OoV words from the calculation, in the experiment on SYOS dataset POST-AL would yield slightly higher accuracy than our SVMTool model (1238 versus 1234 correct predictions).

Test data	Accuracy
TDOA	1910 / 2023 (94.41%)
SYOS	1225 / 1847 (66.32%)

Table 2: Most Frequent Tag baseline.

		Direction (`-S`)		
		LR	**LRL**	**RL**
Tagging strategy (`-T`)	0	97.33%	97.08%	89.77%
	1	-	-	90.46%
	2	97.62%	97.23%	90.21%
	4	97.78%	-	-
	5	97.33%	96.89%	90.11%
	6	**97.83%**	-	-

Table 3: Results (Accuracy) of the experiments with SVMTool on TDOA dataset (best result in bold).

		Direction (`-S`)		
		LR	**LRL**	**RL**
Tagging strategy (`-T`)	0	74.93%	74.07%	69.95%
	1	-	-	69.46%
	2	76.99%	76.77%	72.93%
	4	**78.34%**	-	-
	5	75.09%	75.26%	70.06%
	6	78.07%	-	-

Table 4: Results (Accuracy) of the experiments with SVMTool on SYOS dataset (best result in bold).

Test data	Accuracy
TDOA	1939 / 2023 (95.85%)
SYOS	1238 / 1847 (67.03%)

Table 5: Results of the experiments with POST-AL.

Differences in accuracy observed between various tagging strategies offered by SVMTool were also mainly caused by different scores for unknown words, while the results for known words

[4] For details please refer to SVMTool's documentation (Giménez and Márquez 2012).

Category	Accuracy
Known	1950 / 1977 (98.63%)
Unknown	29 / 46 (63.04%)

Table 6: Results of the experiments with SVMTool (`-T 6 -S LR`) on TDOA dataset – Accuracy per category of words.

Category	Accuracy
Known	1234 / 1410 (87.52%)
Unknown	213 / 437 (48.74%)

Table 7: Results of the experiments with SVMTool (`-T 4 -S LR`) on SYOS dataset – Accuracy per category of words.

exhibited much less variance. For instance, in the experiment on SYOS data, the performance for in-vocabulary words was less than 1% higher with the tagging strategy set to `-T 4` as compared to `-T 0` (1234 versus 1224 correct predictions), but at the same time the performance for OoV words improved by over 12% (213 versus 160 correct predictions).

The best performance with both sets of test data was achieved by tagging strategies 4 and 6. According to SVMTool's technical manual (Giménez and Márquez 2012), both of them utilize Model 4 – the variant which addresses the problem of OoV words by artificially marking a portion of the training data as unknown during the learning process. Additionally, tagging strategy 6 maximizes the global (sentence-level) sum of SVM scores, rather than making decisions based on a reduced context.

Contrary to the results reported by Giménez and Márquez (2012), using the combination of both tagging directions (`-S LRL`) did not improve the performance in our experiments – the only case where it yielded slightly higher accuracy than tagging from left to right (`-S LR`) was the experiment on SYOS with tagging strategy set to 5. The reasons behind this behavior shall be investigated in future research.

8.1 Combined accuracy

Apart from using each of the two taggers in isolation, we are also interested in the possibilities of combining them to maximize accuracy. In order to estimate the potential performance of such combination, we calculated to what extent both taggers agree on their output and how accurate those shared predictions are. In the case of SVMTool, we used the predictions with the highest accuracy for each of the two test datasets (i.e. the ones generated with tagging parameters set to `-T 6 -S LR` for TDOA and `-T 4 -S LR`, for SYOS). Results are shown in Table 8.

The accuracy of shared predictions is higher than the total accuracy of either of the two taggers used in isolation. In the future we might leverage this fact to reduce the amount of incorrect annotations when applying both taggers in a cross-training scenario to bootstrap POS annotations for a larger corpus of Ainu texts.

Test data	Common predictions	Shared accuracy
TDOA	1953 / 2023 (96.54%)	1932 / 1953 (98.92%)
SYOS	1317 / 1847 (71.30%)	1196 / 1317 (90.81%)

Table 8: Proportion of common predictions and their accuracy.

9 Conclusions and future work

In this research we used a small amount of part-of-speech annotated Ainu language textual data to train and compare two POS taggers: POST-AL – a system developed specifically for Ainu, based on contextual (n-gram) and statistical (TF) information derived from a lexicon, and a tagger generated using SVMTool – an off-the-shelf generator of sequential taggers based on Support Vector Machines.

Experiments conducted on two different sets of objective data revealed that the SVM based approach is more effective, especially when applied to out-of-domain data, the main reason for higher accuracy being its ability to predict part-of-speech tags for Out of Vocabulary words.

One of the main tasks for the future is to convert other existing Ainu language resources including the information about parts of speech (such as the dictionary by Kirikae (2003)) to a corpus format which could be used with SVMTool or other POS taggers. We also plan to apply POST-AL and SVMTool in a cross-training experiment to bootstrap part-of-speech annotations for a bigger corpus of texts in the Ainu language.

References

Pitambar Behera, Atul Kr. Ojha, and Girish Nath Jha. 2015. *7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27-29, 2015, Revised Selected Papers*, pp. 393-406.

Anna Bugaeva and Shiho Endō (eds.). 2010. A Talking Dictionary of Ainu: A New Version of Kanazawa's Ainu Conversational dictionary. Retrieved November 25, 2015 from http://lah.soas.ac.uk/projects/ainu/

Yukie Chiri. 1923. *Ainu shin-yōshū* [Ainu songs of gods]. Kyōdo Kenkyūsha, Tokyo

Stephen Clark, James R. Curran and Miles Osborne. 2003. Bootstrapping POS taggers using Unlabelled Data. School of Informatics. University of Edinburgh.

Jesús Giménez and Lluís Màrquez. 2004. SVMTool: A general POS tagger generator based on Support Vector Machines. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*.

Jesús Giménez and Lluís Màrquez. 2012. *SVMTool: A general POS tagger generator based on Support Vector Machines. Technical Manual v1.4*. TALP Research Center, LSI Department, Universitat Politécnica de Catalunya.

Tjerk Hagemeijer, Michel Généreux, Iris Hendrickx, Amália Mendes, Abigail Tiny, and Armando Zamora. 2014. The Gulf of Guinea Creole Corpora. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pp. 523-529.

Hokkaidō Utari Kyōkai [Hokkaido Ainu Association]. 1994. Akor Itak [Our Language]. Sapporo.

Kotora Jinbō and Shōzaburō Kanazawa. 1898. *Ainugo kaiwa jiten* [Ainu conversational dictionary]. Kinkōdō Shoseki, Tokyo.

Hideo Kirikae. 2003. *Ainu shin-yōshū jiten: tekisuto, bumpō kaisetsu tsuki* [Lexicon to Yukie Chiri's Ainu shin-yōshū with text and grammatical notes], Daigaku Shorin, Tokyo.

Yoshio Momouchi, Yasunori Azumi, and Yukio Kadoya. 2008. Research Note: Construction and Utilization of Electronic Data for Ainu Shin-yōsyū. *Bulletin of the Faculty of Engineering at Hokkai Gakuen University*, Vol. 35, pp. 159–171.

Hiroshi Nakagawa. 1995. *Ainugo Chitose Hōgen Jiten* [Dictionary of the Chitose dialect of Ainu]. Sōfūkan, Tokyo.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2017. Improving Tokenization, Transcription Normalization and Part-of-speech Tagging of Ainu Language through Merging Multiple Dictionaries. In: *Proceedings of the 8th Language & Technology Conference (LTC'17)*, pp. 317-321.

Karol Nowakowski, Michal Ptaszynski, and Fumito Masui. 2018. A proposal for a unified corpus of the Ainu language. *IPSJ SIG Technical Report*, Vol. 2018-NL-237, pp. 1-6.

Michal Ptaszynski and Yoshio Momouchi. 2012. Part-of-Speech Tagger for Ainu Language Based on Higher Order Hidden Markov Model. *Expert Systems With Applications*, Vol. 39, Issue 14 (2012), pp. 11576-11582.

Kirsten Refsing. 1986. *The Ainu language. The morphology and syntax of the Shizunai dialect.* Aarhus University Press, Aarhus.

Masayoshi Shibatani. 1990. *The languages of Japan*. London: Cambridge University Press.

Suzuko Tamura. 1996. *Ainugo jiten: Saru hōgen. The Ainu-Japanese Dictionary: Saru dialect*. Sōfūkan, Tokyo.

A Feature set used in experiments with the SVMTool

Tables 9-12 present the feature sets defined for each of the four variants (Model 0/1/2/4) of the tagging model created in this research. Each of the tagging strategies offered by the SVMTool utilizes different variant(s) of the tagging model. For details, please refer to Giménez and Márquez (2012).

Feature category		Definition
Word features		w_{-2}, w_{-1}, w_0, w_1, w_2
POS features		p_{-2}, p_{-1}
Ambiguity classes		a_0, a_1, a_2
Maybe's		m_0, m_1, m_2
Word bigrams		(w_{-2}, w_{-1}), (w_{-1}, w_0), (w_0, w_1), (w_{-1}, w_1), (w_1, w_2)
POS bigrams		(p_{-2}, p_{-1}), (p_{-1}, p_1), (p_1, p_2)
Word trigrams		(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_1), (w_{-1}, w_0, w_1), (w_{-1}, w_1, w_2), (w_0, w_1, w_2)
POS trigrams		(p_{-2}, p_{-1}, p_1), (p_{-1}, p_1, p_2)
Only for OoV words	Single characters	ca(1), cz(1)
	Prefixes	a(2), a(3), a(4)
	Suffixes	z(2), z(3), z(4)
	Lexicalized features	L (word length), SA (initial upper case), AA (all upper case), SN (starts with number), CA (any capital letter), CAA (several capital letters), CP (contains a period), CC (contains a comma), CN (contains a number), MW (contains a hyphen)

Table 9: Feature definition for Model 0.

Feature category		Definition
Word features		w_{-2}, w_{-1}, w_0, w_1, w_2
POS features		p_{-2}, p_{-1}, p_1, p_2
Ambiguity classes		a_0, a_1, a_2
Maybe's		m_0, m_1, m_2
Word bigrams		(w_{-2}, w_{-1}), (w_{-1}, w_0), (w_0, w_1), (w_{-1}, w_1), (w_1, w_2)
POS bigrams		(p_{-2}, p_{-1}), (p_{-1}, p_0), (p_{-1}, p_1), (p_0, p_1), (p_1, p_2)
Word trigrams		(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_1), (w_{-1}, w_0, w_1), (w_{-1}, w_1, w_2), (w_0, w_1, w_2)
POS trigrams		(p_{-2}, p_{-1}, p_0), (p_{-2}, p_{-1}, p_1), (p_{-1}, p_0, p_1), (p_{-1}, p_1, p_2)
Only for OoVs	Prefixes	a(1), a(2), a(3), a(4)
	Suffixes	z(1), z(2), z(3), z(4)
	Lexicalized features	L, SA, AA, SN, CA, CAA, CP, CC, CN, MW

Table 10: Feature definition for Model 1.

Feature category		Definition
Word features		w_{-2}, w_{-1}, w_0, w_1, w_2
POS features		p_{-2}, p_{-1}
Ambiguity classes		a_0
Maybe's		m_0
Word bigrams		(w_{-2}, w_{-1}), (w_{-1}, w_0), (w_0, w_1), (w_{-1}, w_1), (w_1, w_2)
POS bigrams		(p_{-2}, p_{-1})
Word trigrams		(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_1), (w_{-1}, w_0, w_1), (w_{-1}, w_1, w_2), (w_0, w_1, w_2)
Only for OoVs	Prefixes	a(1), a(2), a(3), a(4)
	Suffixes	z(1), z(2), z(3), z(4)
	Lexicalized features	L, SA, AA, SN, CA, CAA, CP, CC, CN, MW

Table 11: Feature definition for Model 2.

Feature category		Definition
Word features		w_{-2}, w_{-1}, w_0, w_1, w_2
POS features		p_{-2}, p_{-1}
Ambiguity classes		a_0, a_1, a_2
Maybe's		m_0, m_1, m_2
Word bigrams		(w_{-2}, w_{-1}), (w_{-1}, w_0), (w_0, w_1), (w_{-1}, w_1), (w_1, w_2)
POS bigrams		(p_{-2}, p_{-1}), (p_{-1}, p_1), (p_1, p_2)
Word trigrams		(w_{-2}, w_{-1}, w_0), (w_{-2}, w_{-1}, w_1), (w_{-1}, w_0, w_1), (w_{-1}, w_1, w_2), (w_0, w_1, w_2)
POS trigrams		(p_{-2}, p_{-1}, p_1), (p_{-1}, p_1, p_2)
Only for OoVs	Prefixes	a(1), a(2), a(3), a(4)
	Suffixes	z(1), z(2), z(3), z(4)
	Lexicalized features	L, SA, AA, SN, CA, CAA, CP, CC, CN, MW

Table 12: Feature definition for Model 4.

Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead

Rogier Blokland
Uppsala University
rogier.blokland@moderna.uu.se

Niko Partanen
Institute for the Languages of Finland
niko.partanen@kotus.fi

Michael Rießler
University of Bielefeld
michael.riessler@uni-bielefeld.de

Joshua Wilbur
University of Freiburg
joshua.wilbur@skandinavistik.uni-freiburg.de

Abstract

The systematic integration of pre-digital published transcriptions of legacy language materials offers many possibilities to enrich documentary corpora with data that is often very comparable to contemporary collections, and often originating from the same speech communities researchers currently work with. Especially recent advances in text recognition technologies make the reuse of old materials a very attractive and accessible task. However, the output of text recognition needs to be connected to further parts of the pipeline, namely forced alignment and speech recognition. The workflows discussed here attempt to reach a maximally useful situation where legacy data is transformed into a usable and comparable format, but not yet transformed into a time aligned corpus.

1 Introduction

This paper discusses opportunities for and challenges of an approach in documentary linguistics which systematically integrates previously published, pre-digital heritage data into a corpus. Based on our own experience, we aim to develop better practices and standards for building more significant corpora in the context of endangered language documentation and description, including potentially any available linguistic data beyond our own annotated fieldwork recordings. Corpora are not only needed for empirically sound descriptions of endangered languages, but can also be utilized in various ways in future computational linguistic studies on these languages.

Although in many cases language documentation work starts from scratch, this is not always the case, such as when previous generations of researchers have produced very large recorded and transcribed collections for the same languages, sometimes even with the same language communities or ancestors of the current speakers. Woodbury (2003) mentions the curation of huge tape collections as upcoming work, and since these collections often connect into already transcribed and published versions of the given texts, our approach aligns very closely with this task. The relevant materials may be handwritten, or partially published in print, and the original recordings are usually scattered in various archives and personal collections, possibly forgotten or even lost. Including heritage data in contemporary corpora is not an easy task. Indeed, it can be overwhelming and very challenging, which makes the temptation to work primarily with new, self-collected data very strong. However, we argue that heritage data is important enough that resources should be systematically allocated to including these data in language documentation projects.

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 24–30,
Honolulu, Hawai'i, USA, February 26–27, 2019.

The authors of this paper have worked extensively with Zyrian Komi, and various Saamic languages (all in the Uralic language family). Examples of the publications integrated into our corpora are parts of Yrjö Wichmann's *Syrjänische Volksdichtung* (Komi spoken texts collected in 1880s, published in 1916), T.E. Uotila's *Syrjänische Texte* (Komi spoken texts collected in 1940s, published in 1986–2006) and Erik Vászolyi's *Syrjaenica* series (Komi materials collected in 1960s, published i.e. in 1999), Arvid Genetz' *Sprachproben* (Akkala, Kildin, Skolt, and Ter Saami spoken texts collected in the 1870s, published in 1891), Georgi Kert's *Obrazcy saamskoj reči* (Kildin and Ter Saami spoken texts collected in the 1950s and 1960s, published in 1961), as well as Ignácz Halász' Pite Saami text collections published in 1893, Eliel Lagercrantz' Pite Saami texts from 1921 (published in 1957 and 1963) and numerous archived materials collected by Israel Ruong throughout his carrier.

Whereas the oldest materials, such as those by Wichmann and Genetz, are already in the Public Domain, the reuse of newer materials had to be negotiated with different stakeholders if the data has not been openly licensed by the publisher already. Methods for accessing these materials have included manual retyping, retrieving text from original digital files and building new OCR models for digitization. Although there may be a time and place for such approaches, this paper emphasizes the most automatized methods, with the wish to streamline the process even further.

Our discussion focuses on text collections published for scientific use, typically as aligned transcriptions and translations in a monograph. This is somewhat distinct from the needs that arise around the use of other community-created resources, such as literature and other truly written-mode texts. Another topic that we do not discuss here is the digitization of dictionaries (as addressed e.g. by Maxwell and Bills (2017)). We do not focus on specific technical implementations, as these change quickly, but discuss the topic on a more conceptual level. However, our technical pipeline has been documented in a GitHub repository[1] and is openly available.

2 Methodological background

One distinct feature of our work, compared to common methodology in fieldwork-based language documentation projects, has been the continual application of language technology in corpus annotation (Blokland et al., 2015; Gerstenberger et al., 2016, 2017). Using computational linguistic approaches for more automated corpus annotation, a component of Documentary Linguistics which was not mentioned by Himmelmann (1998), has resulted in relatively large corpora (measured in the number of morphosyntactically tagged tokens). Furthermore, we consistently integrate all available legacy data, in addition to our own fieldwork recordings. This is possible because the endangered Northern Eurasian languages we work on have a long research tradition and possess a number of extant textual sources in addition to preliminary descriptions. Text collections published by pre-digital language documenters since the late 19th century century are especially interesting for language documentation. The similarity to contemporary language documentation materials may not be immediately obvious, because typically no audio representations of these texts exist (if they predate contemporary recording technology), or audio representations are not available together with the original recording (if the original recording was archived but not catalogued properly or not archived at all). However, the texts correspond to recent transcribed recordings because they represent transcribed spoken linguistic events; furthermore, they are often accompanied by translations into majority languages, just as in modern language documentation projects. The lack of interlinear glossing does not necessarily differentiate these materials from contemporary work, as the need for such annotations is well worth questioning anyway from a documentation perspective when dealing with languages for which basic phonological and morphological descriptions are already available (like for most endangered languages of Northern Eurasia).

Our practices have concentrated around

[1] github.com/langdoc/ocr-pipeline

efforts to digitize these materials and turn them into structured corpora using the quasi-standard ELAN xml-format.[2] We use ELAN even when no recording exists or none has been made available. This is done in order to restrict the data carrier formats of our corpora to a single format (ELAN); due to the technical requirements of ELAN, utterances are symbolically "time-aligned" in each ELAN file, although time-alignment is irrelevant for such exclusively written-format heritage texts.[3] This solution arises primarily out of convenience. For interoperability with audiovisual materials in documentary corpora and in order to query the whole corpus effectively, we want a systematic structure across all corpus files. Our projects have evolved from fieldwork-based language documentation and ELAN is the best-suited tool we have encountered for aligning audio (and video) recordings with annotations. ELAN also allows offline corpus searches[4] and it has become a quasi-standard for archiving language documentation data.

We keep all original transcription systems as separate tiers in the resulting corpus, while using the primary transcription tier in the same orthographic representation relevant for each of our contemporary documentation projects. As long as the interpretation of the corresponding phonological system is similar, the transliteration from one system to another is relatively easy; however, doing this consistently and most reliably is still a question that needs more attention. With legacy data it is not unusual for each publication to use a different transcription system, so many conversion patterns are needed. Our projects use Git to ensure version control of the ELAN files; this solves the issue to some extent, but still needs additional conventions to keep track of the actual modifications. All computational methods we have used to transliterate between writing systems have been entirely rule-based, and we are currently developing a Finite State Transducer for this.

[2] tla.mpi.nl/tools/tla-tools/elan/

[3] If audio becomes available, it can be added and the annotations aligned later.

[4] Note however that the corpus search capabilities of the program have some questions that need to be addressed, see Wilbur 2019

3 Recent advances in text recognition

Whether there are enough digital texts available or whether Optical Character Recognition (OCR) tools are needed extensively varies from case to case (cf. Arppe et al., 2016, 5). Working with Uralic languages, publications since the 1980s can occasionally be found as original digital files, which, although coming with a myriad of other problems, are usually the easiest source for text retrieval. In cases where the text collections were typeset by hand or when the original text files were lost, the only solutions are retyping the text or performing text recognition.

Thus far it has been challenging to carry out good quality OCR on complex scripts with a large number of diacritics, although we have had minor success with commercial software such as Abbyy FineReader, a solution which comes with additional issues (Partanen, 2017). Good results with open source software have previously been tied to the availability of matching fonts, which used to be a great impediment. In recent years, open source OCR systems such as Tesseract[5] and Ocropy[6] have shifted towards neural network approaches that process individual lines instead of characters; this streamlines the training process. It is still difficult to train OCR models that work consistently across texts from a variety of sources, yet training a model for the transcription used in a single publication or one writing system seems to be doable with surprisingly little effort. In the tests done by Partanen and Rießler (2019), a few hundred lines of manually created training data were enough to bootstrap a useful OCR system.

Many text collections for endangered languages can be considered representative of a complex but very narrow domain for text recognition purposes, but here the fact that machine learning methods tend to excel in tasks with such conditions is a clear advantage. If a specific transcription system is used only in one publication series, there is no need to recognize this writing anywhere outside that specific publication, so we can easily train an OCR system that only works within this con-

[5] github.com/tesseract-ocr/tesseract

[6] github.com/tmbdev/ocropy

text. The aforementioned study by Partanen and Rießler (2019) tried this approach successfully with texts on different languages which used the same writing system. The results indicate that it is possible to train multilingual OCR system as long as the typeface is identical and all characters are within the training data. These approaches could well be extended to publications using the International Phonetic Alphabet, the Uralic Phonetic Alphabet or others.

On a related note, methods for Handwritten Text Recognition (HTR) have also been rapidly improving (Kahle et al., 2017). Instead of being font- or typeface-specific, HTR systems generally learn a specific person's handwriting style with considerable accuracy. However, HTR currently needs more training data than OCR (hundreds of pages to achieve ideal results), and this restricts its application to situations where the necessary data exists. In order to test HTR tools with legacy transcriptions, the Institute for the Languages of Finland (KOTUS) is currently carrying out a series of experiments using the Transkribus platform[7] to recognize handwritten transcriptions of dialectal Finnish. In this case there are approximately 17,000 transcribed pages produced by one person, Eeva Yli-Luukko, between the 1960s and 1980s. A few hundred pages are now aligned line-by-line, and this seems to be enough to reach a recognition accuracy higher than 90%. M.A. Castrén's handwritten notes on endangered Siberian languages from the 19th century run to over 10,000 pages, and preliminary experiments carried out with this material at KOTUS have reached a recognition accuracy of 75%. These numbers are still far from the results achieved with state-of-the-art OCR systems (which reach over 99% accuracy), but further development and fast progress on this front is sure to happen. The HTR model training is based on material annotated manually in the Manuscripta Castréniana project, which also publishes digital editions.[8]

[7] transkribus.eu
[8] sgr.fi/manuscripta/

4 Towards structured data

In a typical language documentation corpus, the transcribed utterances are time-aligned to the recorded audio. The metadata contain additional information about the speech event and participants. Since the audio represents the primary data on which the transcription is based, the ELAN file itself stores the links between audio and transcription. In some situations with legacy texts, the recorded audio does not exist, and then the representation on the page can be considered the closest we have to the primary data. This raises the question whether instead of audio links we would need to store information about the sentence's coordinates on the page. Modern OCR software has XML export formats that contain this information, and technically it is possible to connect the coordinates to utterance references. This may seem unnecessary when the text comes from a more contemporary publication, but the older and rarer the source, the more essential this seems. This information concerning the location on the page is essential if one wants, for instance, to link an image to the page with highlighting, or to connect annotations to a digital facsimile of the publication. This linkage becomes more complex when there are both audio and printed images representing the same speech event, although technically the utterance specific metadata could be just enhanced with the time codes. However, this becomes even more complicated when the text and the audio deviate from one another, and the transcription is manually corrected and edited. In this setting, both what would be normally transcribed in ELAN and what has been published in the text collection are derivatives of the original audio. The links between these versions would be useful for various purposes, but this becomes difficult when the utterances are eventually edited and corrected in the ELAN version which has the audio link. Published versions are often edited in various ways, which leads to a mismatch between the published text and the original speech event.

Written text may contain inherent structures, such as information about who is speaking which utterance, and what the consecutive order of the utterances is. Usually no refer-

ence to time codes is made, so the text and audio have to be aligned separately. Forced alignment tools have been tested with language documentation data and even suggested for use with legacy data (Strunk et al., 2014, 3942). However, in our experience, the lack of exact correspondence between audio and transcription, combined with the frequent overlapping speech common in spoken language, prevents current forced aligners from performing sufficiently. Because this work – at least theoretically – should someday be possible, we have not engaged in extensive manual alignment, but are waiting for improved automatic tools to be developed.

Splitting utterances or word-length segments into words and phonemes already works with significant accuracy (cf. Kempton, 2017). Language documentation projects often produce enough transcriptions aligned with audio that some sort of a customized forced aligner could be trained using this data, as has been tested recently with Tongan documentary data (Johnson et al., 2018). Similarly, the experiments with speech recognition for endangered languages discussed in Foley et al. (2018); Adams et al. (2018) offer some new possibilities for forced alignment as well, since erroneously recognized speech could probably be matched with more correct textual transcriptions.

5 Wider perspectives and connections

Older publications that contain text collections can easily be found in bibliographical databases, but with archival records, it is more complicated. Archive identifiers can exist for transcriptions in older publications, although sometimes there is only a note indicating the materials exist. We have had occasional success finding recordings described in text collections through harvested archive metadata, but not all archives release their metadata this way (Thieberger, 2016); indeed, in our experience, even when metadata exists it may not be sufficient to identify which recording matches which text.

When the recordings can be located, acquiring the original recordings has been fairly straightforward, and ensuring more wide-ranging usage rights for integrating this material into a corpus can be negotiated together with the archive and the publisher. The majority of material we have worked with was originally collected by researchers who are now deceased, but in some cases the original author may still be alive and interested in participating in digitization efforts. However, note that the archives have traditionally obtained full rights to redistribute the material upon receiving the original recordings. That said, the copyright and ownership questions concerning this kind of resources can be complicated. Unarchived collections are likely the greatest source of difficulties in this respect.

The large number of texts contained in these published text collections alone is reason enough to integrate them without exception into language documentation materials. There are numerous benefits to having as large a corpus as possible, not least from the point of view of language technology. However, there are reasons that go beyond the simple utility of having a large corpus, the most significant of which is more human and connected to working with the speech communities. Typically these materials have been collected by the relatives and ancestors of community members, and sometimes even include speakers still alive today. It is not uncommon for these small print publications to have never been made accessible in the regions they originate from. There are also clear scientific interests in using such materials, and they can be used to plan future documentation work.

With all the languages we have worked with, some of the oldest materials used are in Public Domain, and one possibility which we have been investigating is the publication of the annotated portions of the corpora with very open licenses so that at least some part of the corpus could be used by researchers with no restrictions. This approach has recently been continued by including these resources in treebanks within a Universal Dependencies project (Partanen et al., 2018). However, we are still lacking methodology and practices that would allow us to seamlessly combine distinct research outputs, such as treebanks, into primary materials and archived versions, so that discoverability of all components would be ensured,

even when the data changes and different resources are stored in different repositories: scanned pages in a library's digital system, original audio recordings in one archive, the partly derived language documentation corpus in another archive, and the treebanks in their own repository.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018*.

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of Plains Cree. *CCURL*, pages 1–8.

Rogier Blokland, Ciprian Gerstenberger, Marina Fedina, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2015. Language documentation meets language technology. In Tommi A. Pirinen, Francis M. Tyers, and Trond Trosterud, editors, *First International Workshop on Computational Linguistics for Uralic Languages, 16th January, 2015, Tromsø, Norway*, number 2015:2 in Septentrio Conference Series, pages 8–18. The University Library of Tromsø, Tromsø.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, and David Nash. 2018. Building speech recognition systems for language documentation: The CoEDL endangered language pipeline and inference system (ELPIS). In *6th Internationall Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Arvid Genetz. 1891. *Wörterbuch der Kola-Lappischen Dialekte nebst Sprachproben*. Number 50 in Bidrag till kännedom af Finlands natur och folk. Finska Vetenskaps-Societeten, Helsingfors.

Ciprian Gerstenberger, Niko Partanen, and Michael Rießler. 2017. Instant annotations in ELAN corpora of spoken and written Komi, an endangered language of the Barents Sea region. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 57–66. Association for Computational Linguistics, Honolulu.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Ignácz Halász. 1893. *Népköltési gyűjtemény*, volume 5 of *Svéd-Lapp Nyelv*. Magyar tudományos akadémia, Budapest.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation and Description*, 12:80–123.

Philip Kahle, Sebastian Colutto, Günter Hackl, and Günter Mühlberger. 2017. Transkribus-a service platform for transcription, recognition and retrieval of historical documents. In *International Conference on Document Analysis and Recognition (ICDAR 2017)*, volume 4, pages 19–24.

Timothy Kempton. 2017. Cross-language forced alignment to assist community-based linguistics for low resource languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 165–169, Honolulu. Association for Computational Linguistics.

Georgij M. Kert. 1961. *Obrazcy saamskoj reči. Materialy po jazyku i fol'kloru saamov Kol'skogo poluostrova (kil'dinskij i iokan'gskij dialekty)*. Nauka, Moskva.

Eliel Lagercrantz. 1957. West- und südlappische Texte. Gesammelt und herausgegeben von Eliel Lagercrantz. In (Lagercrantz, 1957–1966).

Eliel Lagercrantz. 1957–1966. *Lappische Volksdichtung*. Number 112,115,117,120,124,126,141 in Mémoires de la Société Finno-ougrienne. Finno-Ugrian Society, Helsinki.

Eliel Lagercrantz. 1963. Texte aus den see-, nord-, west- und südlappischen dialekten: gesammelt, übersetzt und herausgegeben von Eliel Lagercrantz: Index: Verzeichnis der motive und varianten: mythische symbolwelt: Stilkunst und sprache. In (Lagercrantz, 1957–1966).

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91.

Niko Partanen. 2017. Challenges in OCR today. report on experiences from INEL. In *Electronic Writing of RF Peoples: History, Issues, and Perspectives. 16.–17.3.2017, Syktyvkar*, pages 263–273.

Niko Partanen, Rogier Blokland, KyungTae Lim, Thierry Poibeau, and Michael Rießler. 2018. The first Komi-Zyrian Universal Dependencies

treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 126–132. Association for Computational Linguistics, Brussels.

Niko Partanen and Michael Rießler. 2019. An OCR system for the Unified Northern Alphabet. In *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2019)*. Association for Computational Linguistics, Tartu.

Jan Strunk, Florian Schiel, Frank Seifart, et al. 2014. Untrained forced alignment of transcriptions and audio for language documentation corpora using webmaus. In *International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3940–3947, Reykjavík.

Nick Thieberger. 2016. What remains to be done—exposing invisible collections in the other 7,000 languages and why it is a DH enterprise. *Digital Scholarship in the Humanities*, 32(2):423–434.

Toivo Emil Uotila. 1986–2006. *Syrjänische Texte*. Number 186,193,202,221,252 in Mémoires de la Société Finno-ougrienne. Finno-Ugrian Society, Helsinki.

E. Vászolyi-Vasse. 1999. *Syrjaenica. Narratives, Folklore and Folk Poetry from eight dialects of Komi. Upper Izhma, Lower Ob, Kanin Peninsula, Upper Jusva, Middle Inva, Udora*, volume 1 of *Specimina Sibirica*. Seminar für Uralische Philologie der Berzsenyi Hochschule, Szombathely.

Yrjö Wichmann. 1916. *Syrjänische Volksdichtung*, volume 38 of *Mémoires de la Société Finno-ougrienne*. Finno-Ugrian Society, Helsinki.

Joshua Wilbur. 2019. ELAN as a search engine for hierarchically structured, tagged corpora. In *International Workshop on Computational Linguistics for Uralic languages (IWCLUL 2019)*, Tartu. Association for Computational Linguistics.

Anthony C. Woodbury. 2003. Defining documentary linguistics. In Peter K. Austin, editor, *Language Documentation and Description*, volume 1, pages 35–51. SOAS, University of London, London.

A software-driven workflow for the reuse of language documentation data in typological studies

Stephan Druskat and Kilu von Prince
Dept. of German Studies and Linguistics
Humboldt-Universität zu Berlin
Berlin, Germany
{stephan.druskat,kilu.von.prince}@hu-berlin.de

Abstract

Existing language documentation datasets may be reused in typological research projects, if they can be evaluated for suitability. As these datasets may implement the FAIR principles insufficiently, and occur in diverse data formats, data exploration represents an alternative means of evaluation, as well as the core feature of iterative annotation-analysis cycles during the project. This paper presents a semi-automated workflow driven by a set of corpus software, which enables data exploration as part of the research process, and alleviates its cost. The presented software includes a conversion tool to deal with different formats as well as a search and analysis platform for evaluation and exploration. The authors have successfully extended the software, and implemented the presented workflow in a typological research project on the TAM systems of Melanesian languages.

1 Introduction

Corpus-based typological studies on endangered languages rely on corpus data, whose usual sources include language archives, fieldwork, and data exchange between individual researchers. Irrespective of the source, defining and compiling suitable datasets for a typological study – and working with them to answer a research question – is challenging in two respects, as available datasets

1. must be evaluated for reusability in the first place, unless they are produced during fieldwork which targets the research question;
2. may come in different data formats.

Evaluation of the reusability of a language documentation dataset for typological research includes fine-grained assessments, e.g., probing the occurrence of linguistic phenomena in the data. If the data is retrieved from language archives this is ideally supported by detailed metadata attached to the dataset, as well as suitable metadata search functionality provided by the archive web service. If the data is acquired directly from a colleague, the relevant information may also be retrieved from metadata, or through personal communication. For datasets produced during dedicated fieldwork, reusability information for a research question is inherently available. Due to differing metadata models and formats, incomplete metadata, lack of relevant metadata, and lack of suitable search functionality for archives, reusability information may not be retrievable from metadata. In short, datasets may fail to meet the FAIR Data Principles (Wilkinson et al., 2016).[1] In this case, the assessment must be based on in-depth data exploration. Such an exploration across several datasets can be tedious and time-consuming work, especially if the datasets come in different data formats, and hence may require different tools for the exploration process.

The second challenge - having datasets in different formats - can also complicate the actual research workflow, as different tools with different analysis capabilities may be needed for different datasets. Additionally, the analyses provided by different tools may be hard to compare, or even incompatible.

We propose that the challenges presented above can be mastered in a workflow based on a set of corpus linguistics software, *corpus-tools.org* (Druskat et al., 2016). The software set can be used to convert both the dataset candidates that should be evaluated ('dataset candidates') and the actual datasets to be analysed during a study ('research datasets') to a single format, in which they can be imported in the included search and analy-

[1] With respect to metadata, datasets may fail to meet the reusability principle by not providing "meta(data) [that are] richly described with a plurality of accurate and relevant attributes", or "(meta)data [that does not] meet domain-relevant community standards" (Wilkinson et al., 2016, p. 4).

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 31–36,
Honolulu, Hawai'i, USA, February 26–27, 2019.

sis platform, for cross-corpus evaluation and analysis. We present the case study of a typological research project on Melanesian languages, where we have successfully extended and used these tools to evaluate and analyse datasets from different sources and in different formats.

2 A software-driven workflow

The workflow we propose bypasses the obstacles of datasets in different formats and missing metadata through a focus on data exploration, and the application of suitable software, i.e., *corpus-tools.org*. This set of linguistic corpus software includes *ANNIS* (Krause and Zeldes, 2016), an open source search and visualization platform. *ANNIS* is a web application implemented with a Java frontend and exchangeable backends; while older and current versions use the object-relational database management system PostgreSQL[2] as backend, future versions will use the faster custom in-memory graph database *graphANNIS* (Krause, 2019). The software unifies linguistic annotations and structures of corpora in graphs and makes them accessible via its powerful, linguistically-informed query language *AQL*.[3] Results can be displayed in a wide variety of visualizations, including the ones common for linguistic data, e.g., trees, coreference graphs, etc. Further functionality includes automated frequency analysis based on text and annotation as well as structure and subgraph searches, export of results, and the provision of uniquely identifiable references to queries, result sets and single results via generated hyperlinks. Both text and multimedia corpora are supported, and *ANNIS* offers playback of video and audio segments. With *ANNIS* it is possible to conduct cross-corpus data exploration and analysis, independently of original data formats of the corpora.

In order to use the given corpus datasets in *ANNIS*, they must be converted into a format that *ANNIS* can import. To this end, *corpus-tools.org* also includes an open source conversion framework for linguistic data, *Pepper* (Zipser et al., 2011). During a conversion process with *Pepper*, data is mapped to an instance of the meta-model *Salt* (Zipser and Romary, 2010), of which an open source implementation and API in the Java programming language is also a part of *corpus-tools.org*. *Salt* is based on a generic graph with semantically sparse layers that concretize the model and API.[4] Following the mapping to this intermediate model, the data is mapped to the target format. The *Pepper* platform provides the intermediate graph model, an API, a command-line interface, and a plugin mechanism based on OSGi (OSGi Alliance, 2011), a dynamic module system for Java. The mapping process itself is implemented in plugins for import, export, and model manipulation. A *Pepper* workflow description in XML specifies the order and configuration of plugins to be used during conversion.

The described set of software tools make it possible to implement a workflow for the evaluation of candidate datasets, as well as for the actual analysis in typological and other linguistic studies, via cross-corpus search and visualisation, and iterative cycles of annotation and analysis.

The workflow consists of the following steps, see Figure 1.

1. Compilation of candidate datasets
2. Conversion from source formats to *ANNIS* format and import in *ANNIS*
3. Evaluation of candidate datasets' suitability for the study (query, visualization, analysis)
4. Definition of research datasets based on 3
5. Conversion of research datasets to the format of the annotation software (e.g., *MMAX2*, *EXMARaLDA*, *GraphAnno*, *Toolbox*, *TCF*, *Treetagger*, and more)[5]
6. Annotation
7. Conversion from annotation format to *ANNIS* format and import in *ANNIS*
8. Analysis via query/visualization in *ANNIS*
9. Formulation of research results based on 8

Steps 6–8 are usually repeated in iterative cycles of annotation, analysis, adjustment of requirements. One of the main advantages of the proposed workflow is that these steps can be automated by implementing best practices from software engineering: version control and continuous integration (CI, Booch (1992)). The annotated files can be placed in a version control system, which is polled by a CI system.

When changes are committed to version control, the CI system triggers the conversion to

[2] https://www.postgresql.org/

[3] https://github.com/korpling/ANNIS

[4] *Salt* differs from *LAF* (Ide and Romary, 2006) in that: *Salt* allows annotations on edges; *Salt* models relations between tokens and base text where *GrAF* (Ide and Suderman, 2007) uses spans for both; in *Salt*, primary data is part of the model.

[5] For a list of available modules see http://corpus-tools.org/pepper/knownModules.

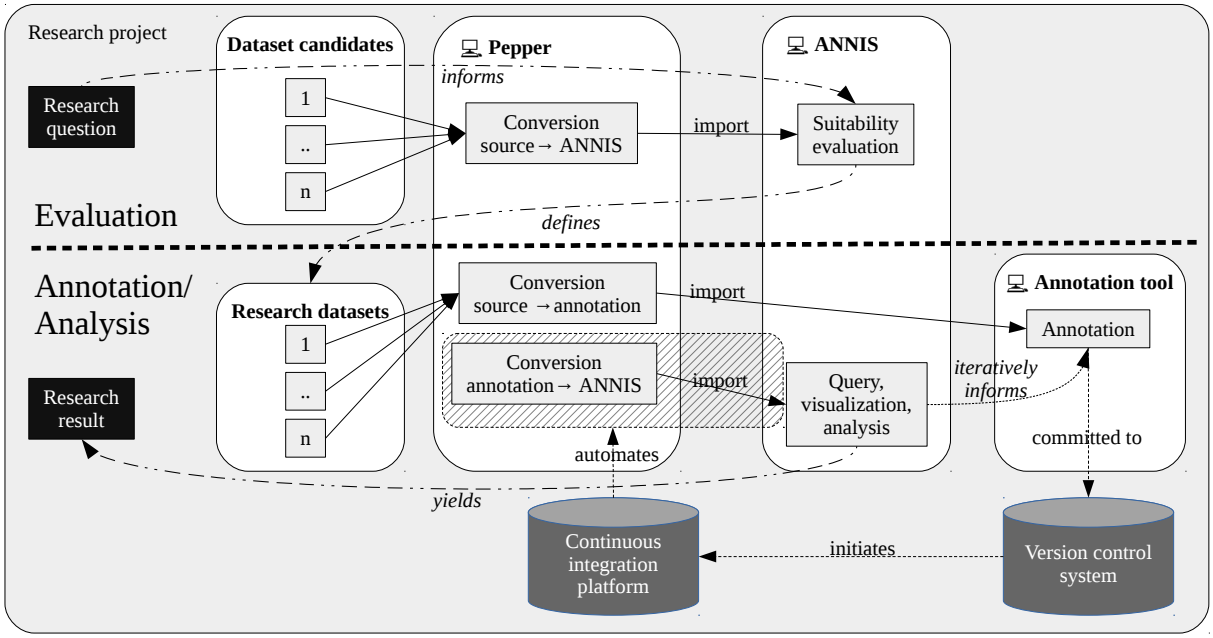

Figure 1: Diagram depicting the proposed workflow based on *corpus-tools.org* software.

the *ANNIS* format via the *Pepper* command-line client, and the subsequent import in *ANNIS* via its REST API. This method also introduces positive side-effects in the workflow: Automated conversion relieves researchers of tedious work; version control allows for unwanted changes to be rolled back; combined with version control, CI introduces reproducibility of annotation-analysis cycles, and enables debugging of erratic processes, and the testability of the automation itself, e.g., by implementing unit and integration tests.

3 Case study: The MelaTAMP project

We have implemented the proposed workflow in the typological research project MelaTAMP.[6]

The project aims to expand the knowledge about tempus, aspect, modality and polarity systems in mood-prominent languages (cf. Bhat (1999)) through a corpus-based analysis of relevant expressions and contexts in 7 Melanesian languages. To this end, the corpora (see Table 1) have been iteratively annotated for clause types, temporal reference, event structure, modal domain, and polarity. Subsequent conversion and visualization in *ANNIS* have enabled the analysis of, e.g., expressions of irrealis (von Prince et al., 2018), or habitual contexts (von Prince et al., accepted).

Within the scope of the project, we have started out with a set of seven existing corpora. The corpora have been acquired through direct exchange with colleagues, and some had been created by project members (see Table 1). We have also added another set of six corpora to the research datasets, that have been elicited as part of the project, using custom storyboards (von Prince, 2018a,b,c,d,e; Krajinović, 2018a,b,c).

Some of the corpora are available from language archives, but the obstacles to data evaluation presented by the lack of relevant metadata and search options in archive interfaces, mentioned in the introduction, pertain, and we have bypassed them by acquiring the respective datasets directly from their authors. Nevertheless, the dataset candidates for our different research questions have been evaluated and selected as intended.

As most of the corpora were available in the *Toolbox* format, we chose to annotate directly in the *Toolbox* text format files with the help of a text editor (Sublime Text 3[7]).

We have automated the annotation-analysis cycle by versioning our *Toolbox* text format files with *Git* (Chacon and Straub, 2014) on our institutional *GitLab*[8] instance. *GitLab*'s CI system has been configured to poll the repository holding the annotations, and run a script on a virtual machine whenever files have changed. The script installs *Pepper*, installs the necessary conversion plugins,

[6] https://hu.berlin/melatamp

[7] https://www.sublimetext.com/

[8] https://about.gitlab.com/

Language	ISO 639-3	Tokens	Country	Elicitator	Format (Software)
Daakie	ptv	~86k	Vanuatu	Krifka (2013)	Text (Toolbox)
Daakie	ptv	~3k	Vanuatu	Manfred Krifa	Text (Toolbox)
Daakaka	bpa	~59k	Vanuatu	von Prince (2013a)	Text (Toolbox)
Daakaka	bpa	~80k	Vanuatu	Kilu von Prince	XML (ELAN)
Dalkalaen		~30k	Vanuatu	von Prince (2013b)	Text (Toolbox)
Dalkalaen		~13k	Vanuatu	Kilu von Prince	XML (ELAN)
North Ambrym	mmg	~24k	Vanuatu	Franjieh (2013)	XML (FLEx)
North Ambrym	mmg	~15k	Vanuatu	Michael Franjieh	XML (ELAN)
Mavea	mkv	~30k	Vanuatu	Guérin (2006)	Text (Toolbox)
Mavea	mkv	~12k	Vanuatu	Valérie Guérin	Text (Toolbox)
South Efate	erk	~54k	Vanuatu	Thieberger (2006)	Text (Toolbox)
South Efate	erk	~15k	Vanuatu	Ana Krajinovic	XML (ELAN)
Saliba/Logea	sbe	~138k	Papua New Guinea	Margetts et al. (2017)	Text (Toolbox)

Table 1: Overview of corpora used in the MelaTAMP project. **Bold** language names signify that a corpus has been elicited during the project. Software: "Toolbox" = (unknown) version of SIL Toolbox (Robinson et al., 2007); "ELAN" = (unknown) version of ELAN (Wittenburg et al., 2006); "FLEx" = (unknown) version of SIL FieldWorks Language Explorer (`https://github.com/sillsdev/FieldWorks`).

converts the annotation data into the *ANNIS* format, and uploads the converted files to an *ANNIS* instance via REST API.

Following the workflow, we needed the following *Pepper* plugins:

- *Toolbox* text format import plugin
- *FLEx* XML import plugin
- *ELAN* import plugin
- *Toolbox* text format export plugin
- *ANNIS* format export plugin

An export plugin for the *ANNIS* format already exists,[9] as does an *ELAN* import plugin.[10] However, the *ELAN* import plugin has been developed for a relatively narrow set of use cases, and did not yield usable results for our case. Instead, we exported the corpus files that were available in the *ELAN* XML format to the *Toolbox* text format via the respective export functionality in *ELAN*.

In the course of the project, we have developed three further plugins in two open source software projects: *ToolboxTextModules* and *FLExModules*.

3.1 ToolboxTextModules

The *ToolboxTextModules* (Druskat, 2018b) project holds a *Pepper* import plugin to map the *Toolbox* text format to a *Salt* model, and an export plugin to map a *Salt* model to the *Toolbox* text format.

[9] `https://github.com/korpling/pepperModules-ANNISModules`

[10] `https://github.com/korpling/pepperModules-ElanModules`

The import plugin is passed a *Toolbox* text format file or a directory containing such files. It will parse the files, validate them, and transform their contents into a *Salt* graph structure of corpora and documents. Documents have their own *document graph*, which contains the language data as nodes and edges. It will contain two base text nodes, whose text values represent:

- the lexical information from lines in the *Toolbox* file marked with `\tx`;
- the morphological information from lines marked with `\mb`.

Token nodes segment the text base according to *Toolbox*' interlinearization; *Salt* can handle *Toolbox*' double segmentation by aligning tokens through a sequential data structure. The plugin also detects invalid interlinearizations in the *Toolbox* data based on incongruencies over token indices, and records them.

Token nodes can have multiple annotations – annotations are realized in the graph model as labels on nodes – and can be covered by span nodes which in turn can have multiple annotations themselves. Span nodes are used to represent `\ref` phrases and `\id` documents.

The import plugin can be passed parameters to configure the conversion. These include: specifications for marker distinction; how morphology delimiters should be handled; whether markers should be changed during conversion; if and how interlinearization errors should be recorded, and

more. Additionally, the plugin supports a custom structure which cannot be expressed in the *Toolbox* software, but is supported by the format and leveraged by our annotation process: Annotations in *Toolbox* can only be assigned to either lexical or morphological units, or to the whole `\ref` span. We have introduced sub-phrases termed *subref*, which are index-determined spans which cover a subset of the complete set of morphological token nodes within a *Toolbox* `\ref` unit. These can be used to annotate, e.g., clauses.

The export plugin converts a *Salt* model to *Toolbox* text format files, and adheres as closely as possible to the Multi-Dictionary Formatter specifications (Coward and Grimes, 2003). Parameters are again used to configure the conversion, mainly to specify which annotation layers in the *Salt* model should be mapped to which markers.

3.2 FLExModules

FLExModules (Druskat, 2018a) provides an import plugin for the XML format exported from *FieldWorks Language Explorer* (*FLEx*). It transforms the interlinearized text structure provided in the XML to a *Salt* model, complete with corpus structure and annotations. This model is very similar to the model produced during conversion from the *Toolbox* text format, but the implementation is less complex than for the import plugin for the latter, given the structured nature of the data. The *FLEx* XML import plugin provides parametrization inasmuch as annotation namespaces and names can be changed in the process using a pre-defined mapping.

4 Conclusion

In the MelaTAMP research project we have successfully implemented and automated the annotation-analysis part of the proposed workflow (see Figure 1) for 13 language documentation corpora provided in three different data formats, and have enabled future implementations of the evaluation part by creating *ToolboxTextModules* and *FLExModules*. In this paper we have also shown that a workflow driven by extended *corpus-tools.org* software can bypass obstacles to data evaluation presented by insufficient implementation of the FAIR principles for data management in language documentation datasets as available in, e.g., archives of endangered languages. The presented workflow features automation as a means to boost efficiency and reduce errors. It enabled us to analyse expressions of irrealis and habitual contexts across 7 Melanesian languages, and subsequently formulate research results such as von Prince et al. (2018) and von Prince et al. (accepted).

5 Acknowledgments

We would like to thank Thomas Krause, the project lead for *ANNIS*, for his continued support. The project "A corpus-based contrastive study tense, aspect, modality and polarity (TAMP) in Austronesian languages of Melanesia (MelaTAMP)" has been funded by Deutsche Forschungsgemeinschaft (DFG) under grant no. 273640553.

References

D.N.S. Bhat. 1999. *The Prominence of Tense, Aspect and Mood*, volume 49 of *Studies in Language Companion Series*. John Benjamins Publishing Company, Amsterdam.

Grady Booch. 1992. *Object-Oriented Design: With Applications*. Benjamin/Cummings, Redwood City, Calif. OCLC: 258275520.

Scott Chacon and Ben Straub. 2014. *Pro Git*, 2nd edition. Apress.

David Coward and Charles E. Grimes. 2003. Making dictionaries: A guide to lexicography and the Multi-Dictionary Formatter.

Stephan Druskat. 2018a. *FLExModules*. DOI: https://doi.org/10.5281/zenodo.1297385.

Stephan Druskat. 2018b. *ToolboxTextModules*. DOI: https://doi.org/10.5281/zenodo.1162207.

Stephan Druskat, Volker Gast, Thomas Krause, and Florian Zipser. 2016. corpus-tools.org: An Interoperable Generic Software Tool Set for Multi-layer Linguistic Corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4492–4499, Portorož, Slovenia. European Language Resources Association (ELRA).

Michael Franjieh. 2013. *A documentation of North Ambrym, a language of Vanuatu*. SOAS, Endangered Languages Archive. https://elar.soas.ac.uk/Collection/MPI67426. [Accessed on 2017/10/04], London.

Valérie Guérin. 2006. *Documentation of Mavea*. SOAS, Endangered Languages

Archive. https://elar.soas.ac.uk/Collection/MPI67426. [Accessed on 2017/03/01], London.

Nancy Ide and Laurent Romary. 2006. Representing Linguistic Corpora and Their Annotations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Nancy Ide and Keith Suderman. 2007. *GrAF: A Graph-Based Format for Linguistic Annotations*.

Ana Krajinović. 2018a. Garden (MelaTAMP storyboards).

Ana Krajinović. 2018b. Haircuts (MelaTAMP storyboards).

Ana Krajinović. 2018c. Making laplap (MelaTAMP storyboards).

Thomas Krause. 2019. *ANNIS: A graph-based query system for deeply annotated text corpora*. Ph.D. thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät.

Thomas Krause and Amir Zeldes. 2016. ANNIS3: A new architecture for generic corpus query and visualization. *Digital Scholarship in the Humanities*, 31(1):118–139.

Manfred Krifka. 2013. *Daakie, The Language Archive*. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.

Anna Margetts, Andrew Margetts, and Carmen Dawuda. 2017. *Saliba/Logea*. The Language Archive. http://dobes.mpi.nl/projects/saliba.

OSGi Alliance. 2011. OSGi Service Platform Core Specification, Release 4, Version 4.3.

Kilu von Prince. 2013a. *Daakaka, The Language Archive*. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.

Kilu von Prince. 2013b. *Dalkalaen, The Language Archive*. MPI for Psycholinguistics. https://hdl.handle.net/1839/00-0000-0000-000F-4E20-B@view, Nijmegen.

Stuart Robinson, Greg Aumann, and Steven Bird. 2007. Managing Fieldwork Data with Toolbox and the Natural Language Toolkit. *Language Documentation and Conservation*, 1(1):44–57.

Nick Thieberger. 2006. *Dictionary and texts in South Efate*. Digital collection managed by PARADISEC. DOI: https://doi.org/10.4225/72/56FA0C5A7C98F.

Kilu von Prince. 2018a. Bananas (MelaTAMP storyboards).

Kilu von Prince. 2018b. Fat pig (MelaTAMP storyboards).

Kilu von Prince. 2018c. Festival (MelaTAMP storyboards).

Kilu von Prince. 2018d. Red Yam (MelaTAMP storyboards).

Kilu von Prince. 2018e. Tomato and Pumpkin (MelaTAMP storyboards).

Kilu von Prince, Ana Krajinović, Manfred Krifka, Valérie Guérin, and Michael Franjieh. 2018. Mapping Irreality: Storyboards for Eliciting TAM contexts. In *Proceedings of Linguistic Evidence 2018*, Tübingen, Germany.

Kilu von Prince, Ana Krajinović, Anna Margetts, Valérie Guérin, and Nick Thieberger. accepted. Habituality in four Oceanic languages of Melanesia. *Language Typology and Universals*.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alexander Klassmann, and Han Sloetjes. 2006. ELAN: A Professional Framework for Multimodality Research. In *Proceedings of Language Resource and Evaluation 2006*, pages 1557–1559.

Florian Zipser and Laurent Romary. 2010. A model oriented approach to the mapping of annotation formats using standards. In *Proceedings of the Workshop on Language Resource and Language Technology Standards*.

Florian Zipser, Amir Zeldes, Julia Ritz, Laurent Romary, and Ulf Leser. 2011. Pepper: Handling a multiverse of formats.

Bootstrapping a Neural Morphological Generator from Morphological Analyzer Output for Inuktitut

Jeffrey Micher
US Army Research Laboratory
2800 Powder Mill Road
Adelphi, MD 20783
jeffrey.c.micher.civ@mail.mil

Abstract

We present a method for building a morphological generator from the output of an existing analyzer for Inuktitut, in the absence of a two-way finite state transducer which would normally provide this functionality. We make use of a sequence to sequence neural network which "translates" underlying Inuktitut morpheme sequences into surface character sequences. The neural network uses only the previous and the following morphemes as context. We report a morpheme accuracy of approximately 86%. We are able to increase this accuracy slightly by passing deep morphemes directly to output for unknown morphemes. We do not see significant improvement when increasing training data set size, and postulate possible causes for this.

1 Introduction

Morphological generation is the process of transforming an underlying sequence of morphemes (for example, a lemma or stem, plus inflections) into a surface realization of those morphemes. For many languages which have complex morphology, morphological generation is a necessary step in certain language processing applications such as machine translation. Finite state transducers (FSTs) have been the main technology used in morphological analysis and generation. Most finite state transducers can operate in both directions, the analysis, or upward, direction and the generation, or downward, direction (Beesley & Karttunen, 2003). The Uqailaut morphological analyzer (Farley, 2009), which provides an analysis of Inuktitut words, however, only operates in the "analyze" direction, because this analyzer was hard-coded, not using currently widely used tools such as *xfst* (Beesley & Karttunen, 2003). To make up for this lack of functionality, we present a method for bootstrapping the output of the morphological analyzer to build a morphological generator that will work in the opposite direction, inspired by the analyzer bootstrapping technique in Micher (2017). We do this using a sequence to sequence neural network architecture based on encoder-decoder networks to "translate" from the deep morpheme sequences to their corresponding surface forms, and we present accuracy results. We show that this technique has promise when building morphological generators when the associated analyzer does not operate in reverse. We envision the use of this generator as a post-process to an English-Inuktitut machine translation system which translates English into sequences of underlying Inuktitut morphemes, to convert those deep morphemes to surface forms.

2 Inuktitut Morphophonemics

Inuktitut is a polysynthetic language spoken in the Canadian territory of Nunavut and other regions of Arctic Canada. Inuktitut words tend to be very long because many morphemes can be added onto roots iteratively, often generating words which would correspond to full clauses in other languages. In general, Inuktitut words consist of a root followed by zero or more lexical postbases, followed by a grammatical suffix and an optional clitic (Dorais, 1990). The surface realization of each morpheme is based on specific morphophomemic rules which are unique to each morpheme and not conditioned wholly on their phonetic environment. For example, the underlying morpheme sequence for the word "mivviliarumalauqturuuq" meaning "he said he wanted to go to the landing strip" is "mit+vik+liaq+juma+lauq+juq+guuq." The spelling rule for each morpheme must be learned individually, and also the final surface spelling

Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 37–41,
Honolulu, Hawai'i, USA, February 26–27, 2019.

will be affected by the previous and following morphemes. We work from the end to the beginning to understand this phenomenon in this example. The morpheme ‘guuq’ is a UVULAR ALTERNATOR [1], which means that the first phoneme ‘g’ will change according to what the previous morpheme ends with. In this case, it surfaces as ‘r’ because of the uvular ‘q’ before it. The rule also deletes a previous consonant, so the ‘q’ of ‘juq’ gets deleted. Next, ‘juq’ is a CONSONANT ALTERNATOR, which means the first consonant gets spelled based on the ending of the previous morpheme. In this case, it comes out as ‘t’ because the previous morpheme ends with a consonant. Next, ‘lauq’ is NEUTRAL, so there are no changes (but if the following morpheme was a UVULAR ALTERNATOR, it would have lost its final ‘q’). Next, ‘juma’ is like ‘guuq’ so it gets spelled ‘ruma’ and deletes the preceding ‘q’. Next, ‘liaq’ is a DELETER, so it deletes the previous morpheme’s final consonant ‘k’ and recall its ‘q’ was already deleted. Next, ‘vik’ is a VOICER, which causes the previous ‘t’ to completely assimilate to ‘v’.

3 The Uqailaut Analyzer and Sample Output

The Uqailaut morphological analyzer takes a single input word and produces an analysis or set of possible analyses for words those words. An analysis consists of a sequence of morphemes in curly braces. Each morpheme consists of its surface form, deep form, and relevant morphological information such as person and number in the case of verbs, as is depicted below:

{<surface form>:<deep form>/<morphological analysis information>}{..}{..}..etc.

The following shows a typical analysis for the word ‘maligarmut,” meaning “bill, law; something that one follows,” in the dative case. For words with ambiguous analyses, each analysis is given on a separate line:

{maligar:maligaq/1n}{mut:mut/tn-dat-s}
{mali:malik/1v}{gar:gaq/1vn}{mut:mut/tn-dat-s}

The Nunavut Hansard data set (Martin, Johnson, Farley, & Maclachlan, 2003), derived from Nunavut legislative proceedings, was processed with the Uqailaut (Micher, 2018a) analyzer to provide data for morphological analysis and machine translation experiments (Micher, 2018b), and is used again in this current set of experiments.

4 Our Model

We consider the task of morphological generation for Inuktitut as a sequence to sequence processing task similar to that of machine translation, and as such, we model our generator after current designs for machine translation and follow the work of (Kann & Schütze, 2017) and (Faruqui, Tsvetkov, Neubig, & Dyer, 2015). Specifically, we use an encoder-decoder architecture with attention (Bahdanau, Cho, & Bengio, 2015) to encode input sequences of morphemes into a hidden state, then decode them into surface characters. The encoder is a bidirectional LSTM (Hochreiter & Schmidhuber, 1997) and the decoder is a character RNN. Different from MT models, however, we limit the encoder to consider only the current, previous, and following morphemes, which we essentially “translate,” letting the attention mechanism figure out that the sequence of three morphemes is focused on the central morpheme. The limiting of the context to the previous and following morpheme reflects the linguistic context for the operation of the morphophonemic rules of Inuktitut discussed above. The figure below depicts the architecture.

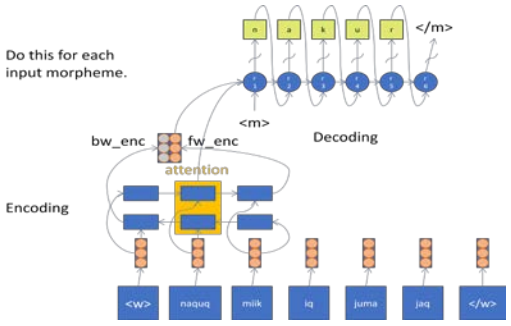

Figure 1: Architecture

[1] The names of the rules are those of (Mallon, 2000)

5 Experiments

For data, we use the same set of morphologically analyzed Inuktitut words as used in (Micher, 2017) having a single morphological analysis, and hold out two sets of 1000 items (labeled 'test' and 'dev'), leaving approximately 23,000 training items.

Our baseline system makes use of the architecture described above. Morpheme embeddings are sized at 128, LSTM hidden states at 128, with 2 layers. We completed 35 epochs with SGD and a simple learning rate. We did not use mini-batching because of the small size of training data. As is common in neural models, we replace infrequent deep morphemes with an <unk> symbol for those having fewer than two instances in the data, and verify that all of the morphemes present in the test sets are found in the training set.

Our second system, Baseline+Backoff, makes use of a simple concept: when an unknown morpheme is encountered, rather than having the system try to generate a correct character sequence, we copy the deep morpheme directly to the surface, with no changes. Because surface forms often contain many of the same character sequences as deep forms, this allows us to "guess" at the right form for those morphemes which may not be sufficiently represented in the data to accurately learn their behavior.

6 Results

We report a full word accuracy score: i.e. the percentage of test items completely correct; a morpheme accuracy score, i.e. how many morphemes are correct, on average, per word; an average Levenshtein edit distance, normalized over word length, by dividing the edit distance score per word by the number of characters in the original word; and a character BLEU-4 (Papineni, Roukos, Ward, & Zhu, 2002) score. While the BLEU-4 score is traditionally used to compare MT system output between different systems, we felt that, as a modified precision score, which accounts for length, we could use it to capture a character-level accuracy. The table below displays the systems and the scores obtained over the results from running the 'test' set through the model.

	Full Word Accuracy	Morpheme Accuracy	Ave. Leven-shtein Distance	Character BLEU-4
Baseline	60.46	86.24	0.036	91.89
Baseline +Backoff	61.50	87.65	0.030	92.64

Table 1: Results

As can be seen, the Baseline+Backoff model performs better on all metrics. It is interesting to note that, even in the non-character based metrics, we get an improvement by simply copying over unknown deep morphemes to the surface. What this reflects is that a certain percentage of morphemes are identical in their deep and surface forms and the model is making mistakes on these, likely, rare morphemes.

We further experimented with adding more training data. From the morphologically analyzed words in the Nunavut Hansard, we used those that had 2, 3, and 4 analyses in addition to the data already used.

# Analyses	# Words	# Training Items
2	28,841	57,682
3	18,519	55,557
4	21,275	85,100

Table 2: Additional Training Data Set Sizes by Number of Analyses

From these data, we created training sets in increments of 25K items over the baseline set, up to 200K items. Each incremental set contained all of the training items from all of the smaller sets, including the baseline data set, using up first the 2-analyses set, then the 3-analyses, and then the 4-analyses sets. As such, the divisions in the training sets do not correspond exactly to the divisions based on number of analyses. We trained using the baseline system described earlier, and tested using the same held out sets as reported for the baseline system. As of the writing of this paper, we are not seeing any significant improvements from the addition of training data over the baseline. All systems are converging at roughly 86% morpheme accuracy, once comparable amounts of data have been seen over several epochs of training. This result may seem counter to general trends in neural network training, in which greater amounts of training data produces better results. However, it should be noted that the different analyses that are provided

by the Uqailaut analyzer are probably noisy: Uqailaut produces all possible analyses whether they are likely or semantically plausible or not. Also, the Uqailaut analyzer is built to account for dialectal spelling variation, which is frequent in the Nunavut Hansard, so learning a single, unambiguous application of a spell out rule of an underlying morpheme sequence from these data may be impossible. A thorough error analysis could shed some light on what is happening with these training data and why additional data are not producing a more accurate system.

7 Future work

Ideally, we would like to refine this approach and get higher accuracy scores, within the 90% and above range, but as accurate as possible since we envision using this model in a downstream machine translation system, in which we hope to minimize the cascading of errors that is often seen in pipelined approaches. Thus, in future work, we will conduct an error analysis to see why there is not more improvement with greater amounts of training data, and we will use an alternative source of data which can be vetted for accuracy and restricted to a single dialectal variant. Also, we will try a comparable system which uses full-word morpheme history instead of only the previous and following morpheme to account for any possible long distance dependencies that may be present in the data. Finally, we will experiment with an unknown morpheme backoff to a character-level encoder, which may show further improvement as specific characters in an unknown morpheme will become salient for purposes of morphophonemic rule application.

8 Related work

Much work has been done on morphological generation. Recent work has focused on morphological reinflection (Cotterell, et al., 2016), (Cotterell, et al., 2017) in which an inflected form is given, and a desired (different) inflected form should be produced. Faruqui et al. (2015) show that a character-level neural model can predict surface forms from base forms plus morphological inflection information. In our work, however, we investigate how well this technique works when only morphological context is provided and no explicit morphological rules or inflection information is given.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *CORR*. Retrieved from http://arxiv.org/abs/1409.0473

Beesley, K. R., & Karttunen, L. (2003). *Finite State Morphology.* Palo Alto: CSLI Publications.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Walther, G., Vylomova, E., Xia, P., . . . Hulden, M. (2017). CoNLL-SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection in 52 Languages. In M. Hulden (Ed.), *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection* (pp. 1-30). Vancouver: Association for Computational Linguistics.

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., & Hulden, M. (2016). The SIGMORPHON 2016 Shared Task—Morphological Reinflection. In M. Elsner, & S. Kuebler (Ed.), *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology* (pp. 10-22). Berlin: Association for Computational Linguistics.

Dorais, L.-J. (1990). The Canadian Inuit and their Language. In D. R. Collins, *Arctic Languages An Awakening* (pp. 185-289). Paris: UNESCO.

Farley, B. (2009). *The Uqailaut Project*. Retrieved from Inuktitut Computing: http://www.inuktitutcomputing.ca/Uqailaut/info.php

Faruqui, M., Tsvetkov, Y., Neubig, G., & Dyer, C. (2015). Morphological Inflection Generation Using Character Sequence to Sequence Learning. Retrieved from http://arxiv.org/abs/1512.06110

Hochreiter, S., & Schmidhuber, J. (1997). Long Short Term Memory. *Neural Computation, 9*(8), 1735-1780.

Kann, K., & Schütze, H. (2017). Unlabeled Data for Morphological Generation With Character-Based Sequence-to-Sequence Models. *CoRR*. Retrieved from http://arxiv.org/abs/1705.06106

Mallon, M. (2000). *Inuktitut Linguistics for Technocrats*. Retrieved from Inuktitut Computing: http://www.inuktitutcomputing.ca/Technocrats/ILFT.php

Martin, J., Johnson, H., Farley, B., & Maclachlan, A. (2003). Aligning and Using an English-Inuktitut Parallel Corpus. Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel

Texts: Data Driven Machine Translation and Beyond - Volume 3 (pp. 115-118). Stroudsburg, PA, USA: Association for Computational Linguistics.

Micher, J. (2017). Improving Coverage of an Inuktitut Morphological Analyzer Using a Segmental Recurrent Neural Network. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 101-106). Honolulu, HI: Association for Computational Linguistics.

Micher, J. (2018a). Provenance and Processing of an Inuktitut-English Parallel Corpus, Part 1. Adelphi, MD: U.S. Army Research Laboratory.

Micher, J. (2018b). Using the Nunavut Hansard Data for Experiments in Morphological Analysis and Machine Translation. *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages* (pp. 65-72). Santa Fe, NM: Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311-318). Stroudsburg, PA, USA: Association for Computational Linguistics.

Author Index