AI for Interlinearization and POS Tagging: Teaching Linguists to Fish

Olga Kriukova[1] *, Katherine Schmirler[2] *, Sarah Moeller[3], Olga Lovick[1], Inge Genee[2], Alexandra Smith[2], Antti Arppe[4]

[1]University of Saskatchewan, [2]University of Lethbridge, [3]University of Florida, [4]University of Alberta
*These authors contributed equally
Correspondence: olga.kriukova@usask.ca

Abstract

This paper describes the process and learning outcomes of a three-day workshop on machine learning basics for documentary linguists. During this workshop, two groups of linguists working with two Indigenous languages of North America, Blackfoot and Dënë Sųłiné, became acquainted with machine learning principles, explored how machine learning can be used in data processing for under-resourced languages and then applied different machine learning methods for automatic morphological interlinearization and parts-of-speech tagging. As a result, participants discovered paths to greater collaboration between computer science and documentary linguistics and reflected on how linguists might be enabled to apply machine learning with less dependence on experts.

1 Introduction

During this time of increased AI-assisted language documentation, more and more studies emphasize the necessity and importance of collaborative efforts between documentary and computational linguists (Gessler, 2022; Flavelle and Lachler, 2023; Opitz et al., 2024). Additionally, Gessler and von der Wense (2024) point out the lack of interdisciplinary educational initiatives that could introduce specialists from both fields to the specific and general context of each other's work and, thus, bring mutual understanding and effective collaboration. In this paper, we describe our experiences hosting and participating in a "Machine-in-the-Loop" (MitL) workshop, held in Edmonton at the University of Alberta during November 14-16, 2023, which addresses this lack. The workshop curriculum Moeller and Arppe (2024) aims to introduce documentary linguists to machine learning (ML) and natural language processing (NLP) and to provide Python-savvy linguists with ML skills relevant to Indigenous language research and resource development. The workshop focused on foundational concepts underpinning machine learning and its application in NLP. In practical sessions, we worked in two teams focusing on two Indigenous languages of North America, Blackfoot and Dënë Sųłiné, each working toward a different project goal using a different machine learning model and NLP task. A Transformer deep learning model was trained to perform automatic interlinear morphological glossing of Blackfoot texts. A Conditional Random Fields (CRF) model was used to build a parts-of-speech (POS) tagger for Dënë Sųłiné.

The paper does not provide any ground-breaking solutions for computational linguistics but rather describes how already-established techniques can facilitate linguists' work with truly under-resourced languages. Notably, the workshop outcomes demonstrate that gaining awareness and a basic understanding of foundational ML concepts, combined with basic programming skills, enables linguists themselves to use NLP for the study and annotation of endangered languages.

This paper advocates for active collaboration between documentary and computational linguists in a way that enables documentary linguists to automate their own work efficiently, thereby reducing reliance on NLP experts to advance language technology for Indigenous communities. We feel that such a collaboration does not happen very often because linguists and computer scientists both assume it takes years of education before one can practically apply machine learning. This, combined with a below-average interdisciplinary dimension in NLP (Wahle et al., 2023), means many attempts at collaboration become inefficient interactions that seem more like data extraction to linguists (Flavelle and Lachler, 2023). This not only raises concerns about data security and sovereignty but also excludes the linguists' and language communities' perspectives from the NLP development. We believe that an approach where collaborators do not assume the technicalities are beyond linguists'

Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 139–149

grasp leads to the effective sharing of knowledge as well as results. For example, we found that, while NLP experts can automate their solutions, documentary linguists can immediately identify the problems in NLP model output, leading to increased problem-solving and benefits to both NLP and documentary goals.

Overall, by describing our workshop experience and our reflections on the interactions, we provide a positive example of collaboration between documentary and computational linguists, showing how much can be achieved in just three days by communicating needs, challenges, problems, and new terminology. We think that such collaborations can benefit both disciplines and support endangered language revitalization and documentation. We take inspiration from the proverb "Give a man a fish and you feed him for a day; teach a man to fish and you feed him for a lifetime" and use the metaphor of teaching a linguist to fish to illustrate the perspectives of both groups in section 2 followed by a description of the workshop in section 3 and the languages in section 4 followed by the outcomes from our three-day "teach a linguist to fish" approach to AI in sections 5 and 6.

2 Perspectives on Machine Learning for Documentary Linguistics

During the MitL workshop, we found ourselves falling into three main groups, each of which has different concerns regarding language data and ML, and thus takes a different approach to "fishing." The first group (co-authors Kriukova and Schmirler) consists of the documentary linguists with an interest in and familiarity with computational methods, who want to actively participate in creating, evaluating, and testing computational models (i.e., learning to fish, see section 2.1). We will refer to them as computationally-minded linguists or CMLinguist1 and 2. The second group (co-authors Genee, Lovick, Smith, to be referred to as DocLinguist1, 2, and 3) is comprised of documentary linguists with less interest in undertaking the computational data processing themselves, but who instead want to be familiar enough with the methods to communicate their needs and evaluate the outputs effectively (i.e., being on board, see section 2.2). The third group (co-authors Arppe and Moeller – to be referred to as CompLinguist1 and 2) contains the computational linguists who organized the workshop and who were primarily concerned with how participants might gain access to sufficiently powerful computing resources and make use of these resources (i.e., the tools used for fishing, see section 2.3). They chose the computational models and code used in the workshop to match the level of technical skills of the linguists in the second group, assuming they would have help after the workshop from those in the first group. Sections 2.1 to 2.3 are written by each group, respectively. Additionally, we want to emphasize that the teaching part of the metaphor is not intended in a strict sense, i.e., we did not expect to turn documentary linguists into computational linguists in a three-day workshop, but rather we aimed to bridge the knowledge gap sufficiently so that documentary linguists could initiate and foster effective collaboration with or without the direct guidance of NLP experts.

2.1 Teach a linguist to fish

Linguists who are primarily trained in language documentation and description, have basic programming skills and have an interest in computational methods are happy to be directly involved in the development of ML applications for endangered languages. We are also interested in working together with computational linguists. However, even when there is an interest, we find barriers to direct involvement or effective collaboration. When we look for help, we find guides for ML online that are either oversimplified or too focused on the mathematical foundations of the algorithms at hand (Vajjala, 2021). Meanwhile, we just want to know enough about ML methods to apply them to our data and to understand how the data and models (or at least the outputs) interact, allowing us to evaluate and improve the models ourselves. Moreover, most guides and books are focused on major languages and thus leave our questions about morphosyntactically different languages unanswered. They are insufficient for beginning work with endangered languages. For example, "Sentiment Analysis Using Python"[1] never mentions "English" but it becomes clear that the guide assumes the language already has a tokenization model, a list of stop words, and morphological model for lemmatization. Also, guides may assume we need the latest and most advanced language models. Sometimes, simple and time-tested methods are sufficient for work with limited data (see section 2.3) and their simplicity and reduced computing demands can

[1] https://www.analyticsvidhya.com/blog/2022/07/sentiment-analysis-using-python

significantly reduce our workload.

Another problem we encounter is that many ML tutorials use pre-built, standardized datasets. Little attention is paid to the description of how to create a dataset for a particular model from scratch (Vajjala, 2021). At the same time, questions such as what file format is needed, what pre-processing is required, or how the metadata file should be organized, are very important to us documentary linguists, who rarely possess "sterile" ready-made datasets.

As computationally-minded linguists who do language documentation work, we are also well-positioned to serve as translators between computational and documentary linguists, who have less or no interest in developing skills in the computational side of our work. This middle-ground understanding allows us to effectively communicate with both computational and documentary linguists about the modeling process and data annotation.

2.2 Get a linguist on board

Documentary linguists with less interest in undertaking computational data processing often take a somewhat ambiguous stance toward NLP. We are interested in utilizing customized NLP tools to manage our data and speed up our analytical work. When relying exclusively on manual annotation by trained individuals, this analytical work is time and labour-intensive, and as a result, can be very expensive as well. We also perceive serious interest in the communities we work with to benefit from the outputs of computational work, in particular in the areas of Automatic Speech Recognition, machine translation, talking dictionaries, and anything that will support the development of pedagogical materials. On the other hand, those communities are concerned about data sovereignty issues with respect to Indigenous language data (see for example, Rainie et al., 2019 or Junker, 2024). We also know from experience that computational linguists routinely underestimate just how "messy" language documentation is at all levels: from noisy multi-speaker audio recordings to code-switching and inconsistent or erroneous transcription and annotation. While we may not be best-suited to learning NLP techniques ourselves, our direct involvement and guiding role in NLP development for the languages we work with has clear benefits for the processes discussed below, particularly as it allows us to act as advocates for the communities who are most likely to suffer any negative impact.

2.3 A fishing rod vs. an industrial trawler

In the context of documentary linguistics, NLP researchers equipped with advanced AI techniques are like industrial fishing trawler operators, aiming to maximize scale, capabilities, and efficiency. We work in a field that often prioritizes publications of cutting-edge performance. However, we find that the needs of documentary linguists which could be served by NLP often require fundamental NLP tasks, such as POS tagging, or are best served by models which are not state-of-the-art. From our experience, linguists' most needed tasks often seem underwhelming. The computational problems involved in documentary work may be viewed in NLP as already solved even in low-resource settings, or the main workload may consist of basic data processing. Therefore, undertaking NLP work to benefit documentary linguistics and minority communities can leave one feeling that we are being asked to leave the trawler and sit on the shore with a bamboo rod.

Yet we found these seemingly mundane tasks are often out of reach for documentary linguists even if their training does include introductory programming skills. They may be unaware of the simplest NLP tools or common low-resource techniques. Designing the workshop we hypothesized that social scientists who discover the regular and irregular structures of language and can describe how they fit in a complex system of previously unstudied languages, all without being able to speak the language, are capable of grasping fundamental concepts of ML. We gambled that linguists could bridge the knowledge gap sufficiently in three days to empower them to design and direct their own collaborative computational projects, even if they could not code one line of Python. We feel the outcomes, whether in a POS tagger, F_1 scores, or the participant's intelligent use of new vocabulary, justified our assumptions. We emphasize that the next steps described in sections 5 and 6 were proposed, explained, and are being independently executed by the linguists themselves.

3 The Workshop

Just as hiring a fishing guide might be advantageous over buying one's own oceangoing trawler, collaboration with NLP experts can be highly beneficial for documentary linguists. However, the advantages of relying on NLP expertise for computationally intensive tasks must be weighed against long-term

dependence on domain experts who have different long-range goals. Also, if the short-term need of the documentary linguist or language community is critical enough to outweigh the downsides of "being given a fish," quick-fix NLP solutions are appropriate. The ideal situation, however, is that linguists themselves would be able to perform the required NLP work. The workshop description below illustrates how this long-term ideal situation can be created in practical terms.

3.1 Summary of the Machine-in-the-Loop Workshop

This workshop aimed to introduce linguists from non-technical backgrounds to the use of NLP for language-related tasks. In the mornings, lectures and interactive activities introduced the participants to the general principles of ML algorithms with clarification of specific relevant topics or terminology such as unsupervised vs. supervised learning and classical machine learning vs deep learning. Special attention was paid to those ML methods that work well in low-resource settings and to precision, recall, and F1 scores for evaluation. In the afternoons, two teams of linguists were able to apply what they learned by training a model on their own data and improving it during the workshop. Discussions and questions were encouraged, as well as sharing progress and roadblocks between the two teams.

3.2 ML for language documentation

While a fuller account of the curriculum of our workshop can be found in Moeller and Arppe (2024), here we briefly summarize our understanding and use of ML in the workshop. We define ML as a type of AI, wherein a computer makes use of an algorithm and statistical model to do something "intelligent". Data are mapped as points in space, and ML creates a statistical model based on that data, which can then be used for prediction. Predictions are made by learning patterns from data. This pattern recognition somewhat mimics how humans learn, and thus the computer can help improve on a manual task.

ML is already used for some linguistic or language-related tasks, such as clustering for dialectology or n-grams for predictive text, but can be also useful for documentary linguistics, particularly in the data annotation bottleneck. Taking audio or transcribed data from its raw form to a fully interlinearized corpus is a time-consuming process. Since linguists already create some transcribed and annotated data as part of their basic analysis, ML offers the opportunity to use those annotations as training data for predictive models to speed annotation for the remaining data.

In the workshop, we take a machine-in-the-loop approach (active learning) that allows human linguistic expertise to annotate new data selected based on the marginal probabilities of a CRF, for example. This approach assists simultaneously in completing the annotation process and more quickly improving the model's output. The workflow involved gathering and preparing our data for training (ideally ahead of time for preprocessing), choosing a model, and then training, testing, and evaluating the model. The linguists evaluated the model, decided what changes to the data were needed, and updated the training dataset.

4 Languages and Data

4.1 Blackfoot

Niitsi'powahsin or Siksikai'powahsin, usually called Blackfoot (ISO: bla) in English, is an Algonquian language spoken in Alberta and Montana by perhaps less than 5,000 people out of a total population of around 40,000 (Genee and Junker, 2018, 301–302). The data used in the workshop is a collection of stories containing ∼1,000 words, drawn from several sources (Russell and Genee, 2014; Ermineskin and Howe, 2005; Genee, 2009; Frantz, 2017; Glenbow Museum, n.d.; Many Feathers et al., 2013), interlinearized as in (1).[2]

(1) Ninna iikaahsitapiiwa.

n-inn-wa
1-father-AN.SG
iik-yaahs-itapii-wa
very-kind-be_person.VAI-3SG

'My father was a very kind person.' (Russell and Genee, 2014, 12)

Blackfoot is a polysynthetic language with many possible morphemes per word. While generally concatenative, these morphemes also display considerable allomorphy and surface variation due to morphophonological processes, which, as we will discuss in our outcomes, can cause issues with data

[2]Most sources provided the analyses for the interlinear glossing. For Russell and Genee (2014), the analyses were provided by DocLinguist1 and for Glenbow Museum (n.d.), the analyses were provided by Heather Bliss, November 2010.

preprocessing for machine learning and no doubt offers an extra challenge for the model itself by introducing considerable variation and ambiguity.

The team working with Blackfoot data consisted of three members. DocLinguist1 is working on Blackfoot language documentation and revitalization and has more than a decade of experience with this language. CMLinguist2 is a postdoctoral scholar who specializes in Algonquian linguistics with a focus on morphosyntactic and phonological modeling and currently works with DocLinguist1. DocLinguist3 is an MA student of DocLinguist1 and a member of the Piikani Nation, who also works in the field of documentation and revitalization of Blackfoot, with a focus on corpus creation and textual annotation.

4.2 Dënë Sųłné

Dënë Sųłné (ISO: chp; hence: DS) is a Dene (Athabaskan) language spoken in Manitoba, Saskatchewan, Alberta, and the Northwest Territories (Cook, 2004). The 2021 census indicates that there are around 10,000 speakers of DS (Statistics Canada, 2022), making it one of Canada's most vital Indigenous languages. More than half of these speakers reside in Northern Saskatchewan. The data used in the workshop is a sub-corpus of the audiovisual corpus compiled during the *Talking Dene* project [3], which collected 70 hours of naturalistic DS representing 100 speakers ranging in age from 13 to 83 years of age. Most of this corpus has been transcribed and translated (at the utterance- and word-level) by speakers fluent in DS and English as shown in (2). The dataset is not made available here due to community preferences.

(2) grade two *dé́ dlą́t'ı së*teacher *nı sí́ bënasní=lë hotıé́ dódí*

grade	*two*	*dé́*	*dlą́t'ı*
grade	two	when	who
së-teacher		*nı*	*sí́*
1SG.PSR-teacher		PST1	EMPH

bë-n-a-s-ní=lë
3SG.P-LX-LX-IPFV:1SG.S:VV-remember=NEG
hotıé́ dódí
very NEGEX

'In grade two, I don't remember who my teacher was at all.' ITN-ETM-2022-11-28-AB

[3] A University of Saskatchewan research project (PI - Olga Lovick) in partnership with the Clearwater River Dene School and the University of Zurich.

From a typological perspective, DS can be described as highly synthetic and fusional, particularly in the verbal domain. The language is overwhelmingly prefixing, with lexical/derivational and inflectional morphemes interspersed within the verb word. It is head-marking and has SOV word order although the fact that arguments are marked on the verb means that full noun phrases are used more sparingly than in languages such as English.

Example (2) illustrates one of the more challenging aspects of our DS corpus: the extensive use of English. Almost all speakers of DS nowadays are bilingual, and switches ranging from one word to multiple sentences, as well as English stems with DS affixes, are extremely common.

The DS team consisted of three individuals. DocLinguist2 is a specialist in Dene/Athabaskan linguistics with over two decades of experience in the description and documentation of this language family. CMLinguist2 is DocLinguist2's and CompLingust1's Ph.D. student. Her research area is in harnessing computational tools for educational and documentary purposes in low-resource language settings, in particular for DS. In addition, we had one graduate student observing the workshop and the work of the DS team, though their own research concerns neither DS nor Blackfoot.

5 Modeling outcomes

5.1 Blackfoot

5.1.1 Process

Our Blackfoot team used the Transformer deep learning model (Vaswani et al., 2017) for the automatic interlinearization of Blackfoot text, specifically morphological segmentation and morpheme glossing. Approximately 10% of the manually annotated data was set aside for testing, with the remaining 90% used for training.

5.1.2 Outcome

For the first training iteration, our Blackfoot team found that a small number of closed-class morphemes achieved promising precision and recall scores. These included the demonstrative stem *ann-* 'that' (0.75 precision and recall, n = 4 in the test dataset) and the demonstrative suffix *-hka* 'invisible' (0.67 precision and recall, n = 6). However, we quickly learned that the glossing in our training data was less consistent than we had thought. Our team thus needed to identify and correct inconsistencies in the glossing. The same morphemes

may have been given slightly different English glosses (e.g., *aakii* 'woman' or 'lady'; *sook-* 'suddenly' or 'unexpected'), allomorphs were separated (e.g., *-nnaan*, *-innaan*, *-(i)nnaan* '1PL', *n-*, *ni-*, *nit-*, *ni(t)-* 'first person'), or different abbreviations were used (e.g., NONSP, NSPEC for 'nonspecific'). Especially with such small datasets, consistency in morphemes and glosses can drastically increase the number of different training examples of each feature.

In the second iteration, many of the morphemes our team corrected showed improvements, such as *aakii* 'woman' (0.67 for both precision and recall, n = 3), demonstrative stems *am-* (0.75 for precision and 1.00 for recall, n = 3), *amo-* (1.00 for precision and 0.67 for recall, n = 3), and *ann-* (precision increased from 0.75 to 1.00, recall unchanged at 0.75, n = 4). The suffix *-hka* also improved (from 0.67 for precision and recall to 0.71 and 0.83 respectively, n = 6), as did the first person prefix *nit-*, now combined with its allomorphs (0.90 for precision and 0.82 for recall, n = 11).

More glossing issues were found and adjusted before the third iteration, but fixing inconsistencies no longer seemed to affect the training, so our team expanded the training data by adding five new sentences, which is a 3.2% increase in training tokens. For our fourth and final iteration of the workshop, some improvements were seen, such as for *ann-* (0.80 precision, 1.00 recall, n = 4) and *nit-* (1.00 precision, 0.50 recall, n = 12). Other affixes also showed promise, such as the third person prefix *ot-* (0.67 precision, 0.50 recall, n = 4), the animate singular suffix *-wa* (0.67 precision, 0.15 recall, n = 13), and the singular suffix for inanimate nouns *-yi* (0.44 for both precision and recall, n = 9).

Overall, our team found that consistent and frequent inflectional morphemes and frequent stems without considerable allomorphy were recognized well. Some demonstrative stems, noun and verb stems, person morphology, and the particle *ki* demonstrated decent precision and recall scores, some were correctly recognized correctly from the beginning and others after glossing was made more consistent. However, much of the inflectional morphology and most stems were still unrecognized or very poorly identified by the model, and much remains to be done before a useful morphological model is available.

5.1.3 Next Steps

In the future, the Blackfoot team plans to tackle two main issues. First, we intend to develop two sets of strict glossing standards across the existing analyzed texts, with a clearly defined correspondence between them. One set will be for linguistic analysis, with each morpheme represented separately and homophonic morphemes are marked differently for nouns and verbs. The other set will be geared toward machine learning, where frequently occurring strings of morphemes can be chunked together and morphemes are marked the same regardless of the word class they attach to. For example, a morpheme string like *-aanaana* can be broken down into *-a* 'direct', *-innaan* '1PL', and *-wa* '3SG', but for the sake of statistical modeling, it may be worthwhile to consider this frequent sequence of morphemes as one unit glossed '1PL>3SG'. For some homophonic person and number suffixes, such as *-wa*, the linguistic analysis will give different glosses depending on whether it attaches to a verb (3SG) or a noun/pronoun (AN.SG). However, for machine learning, one label (e.g. '3SG') may be more effective, especially for a relatively small dataset.

Additionally, our team has an option for generating Blackfoot words with an FST-based morphological model (Kadlec, 2023). With this option, we can generate potentially (hundreds of) thousands of Blackfoot words for inclusion in the training data, increasing the data exponentially. As the token counts in the previous subsection indicate, an increase in data is much needed. In doing this, we intend to explore to what extent synthetically generated paradigms improve this process, e.g., when do we see diminishing returns with increased data.

5.2 Dënë Sųłné

5.2.1 Process

Our DS team undertook the task of parts-of-speech (POS) tagging for DS using the Conditional Random Fields (CRF) model (Lafferty et al., 2001), chosen because of its ability to learn from very small datasets and to demonstrate the role of features of the data for training. The training data consisted of two files comprising 582 DS utterances and 2961 DS words. DocLinguist2 created a controlled parts-of-speech vocabulary in ELAN (ELAN (Version 6.7) [Computer software], 2023) informed by her grammatical research on Dene/Athabaskan languages (Lovick, 2020) and

tailored to DS. ELAN files were manually annotated for POS by an undergraduate student, hand-corrected by DocLinguist2, and exported as FlexText to facilitate further data extraction by CompLinguist2.

The list of POS tags used by our team comprised 18 items. In a sample of 2,817 words taken from dialogue and monologue, nouns and verbs were the most frequent with 521 and 520 tokens, respectively. Particles, adverbs, postpositions and conjunctions were the next most frequent categories with more than 250 tokens each.

5.2.2 Outcome

For our team's first iteration, we achieved an accuracy score of 0.71. Similarly to the Blackfoot team, inconsistent training data annotation was a major source of our model's poor performance in the beginning. This inconsistency was partly due to grammatical differences between DS and English (the language spoken by the undergraduate annotator). Property concept words, for example, are typically adjectives in English but verbs in DS. Other inconsistencies resulted from the fact that some lexical items are polyfunctional and therefore often annotated for the wrong function in context; e.g. *dë́* can function as a postposition 'when, at the time of' (cf. (2) above) or as a clause conjunction 'if' (see also Cook, 2004, 375–380). To simplify the modeling task, we decided to reduce the number of tags, which led to a slight accuracy improvement to 0.73 over several iterations.

For our fourth iteration, we modified the CRF model features. Initially, we used default word feature extraction parameters designed to capture English POS-specific prefixes (e.g., *re-*, *un-*, *mis-*) and suffixes (e.g., *-ed*, *-s/-es*, *-er*. To address the radically different verb morphology of DS, where a verb stem may be preceded by multiple prefixes, we experimented with different numbers of word-initial and -final characters. The settings that gave us the best results captured up to six word-initial and word-final characters. This change in the word-to-features Python function improved the overall accuracy and verb and noun recall value (from 0.84 to 0.87 and from 0.73 to 0.80 respectively).

After the fourth iteration, we found that further feature engineering led to an improvement in the recall for certain parts of speech, at the cost of recall for others. For instance, the recall of verbs improves from 0.87 to 0.91 when we include up to 5 final characters of a word. However, these settings lower the recall of nouns to 0.73, that of postpositions from 0.73 to 0.65, and that of conjunctions from 0.75 to 0.50. This experimentation taught us that we can adjust the model feature parameters to refine the results in specific areas.

Careful examination of the predicted POS for our team's best iteration revealed that a major source of errors was the presence of English and mixed-language lexical items (such as *sëteacher* 'my teacher' in (2)) present within the DS discourse. Tailoring our feature parameters in the CRF to capture DS morphological features caused the model to perform poorly when faced with English or mixed-language words. Consequently, the overall POS tagging performance for DS words is, in fact, higher than the numbers above suggest.

5.2.3 Next Steps

Given the persistence of code-switching and code-mixing in the DS corpus, it appears that the easiest way to improve the accuracy of POS tagging is to add an intermediate step of language identification. The language recognizer could employ a CRF or another non-neural classifier model such as a Support Vector Machine (SVM) to classify each word as DS, English, or Mixed.

POS tagging will then proceed differently depending on the language of each lexical item. English words will be tagged by a pre-trained tool such as spaCy (Honnibal and Montani, 2017). DS and Mixed items will be tagged by our CRF-based tagger. This will also allow our team to evaluate the 'real' accuracy of this tagger. Additionally, in order to facilitate further linguistic data analysis and to improve word search in ELAN, we need to develop a workflow to import the predicted POS tags back into the ELAN with the *lxml* Python package.

6 Learning outcomes

In this section, we move onto the outcomes that we deem of even more interest than the modeling outcomes—the knowledge, understanding, and skills we gained over the course of the workshop and the methods we learned. We call back to section 2 and reflect on these outcomes by each subgroup at the workshop.

6.1 For computationally-minded linguists

It is very significant to us computationally-minded linguists that we not only made a functioning model but also learned how to adapt it to different needs. The CRF model we developed is not perfect and

probably not optimal, but now we have the knowledge sufficient to maintain, modify, and improve it. Moreover, we have a better understanding of how to use our expertise in the languages at hand for feature engineering. As a result, after the workshop, the DS team trained several CRF-based models for different annotation needs.

Working with a Transformer model for interlinearization gave us a better idea of how the annotation may need to differ between computational and documentary linguistics. Though the computational FST modeling of Blackfoot had already demonstrated this to some degree, the chunking and standardization of morphemes and tags became even more apparent when training a Transformer model on a small data set, and will inform both documentation and computational modeling of Blackfoot in the future.

Finally, we realized that all we needed to launch our independent work with ML-based tools was guidance appropriate to our skill level and field of application and a gentle push in the right direction to use our data for our goals.

6.2 For documentary linguists

From the perspective of documentary linguists without programming skills, an important advantage of the workshop approach is the establishment of trust relationships. By forming small teams including both documentary and computational linguists, we were able to ensure that the data did not leave the servers approved by our community partners and University Research Ethics Boards, which protects the data from unauthorized use. We could also see and control what happened with the data because we were in the same space.[4] We think this should be expanded in future workshops and collaboration to include an even more important relationship: that between language communities and the academic community. Including representation from the language communities will foster transparency and create confidence in the process.

A second advantage of the workshop approach lies in the ability to jointly and immediately look at model output and identify problems. The computational linguist may look at numerical indicators of model results, but only someone intimately familiar with the language under analysis can determine that a particular set of errors is perhaps due to inconsistent glossing within the training data. What's more, we can immediately correct some of the errors or suggest improvements to the model based on our understanding of language's fundamental principles. This effect is maximized by goal-setting and preparation in advance of the workshop (i.e., preparation of training data by linguists).

The ultimate strength of the workshop format is that it allows all participants to bring their unique expertise to the table. Rather than force a computational linguist to clean a dataset, or a language documentation specialist to use the command line, we all perform those tasks that we are best suited to. Documentary linguists do not necessarily want to learn to fish ourselves—we want to see that the boat is going in the right direction and to help you know what fish are worth fishing for.

We may lack the time or inclination to learn how to apply NLP ourselves. However, a basic introduction to NLP concepts enables us to communicate our needs effectively and evaluate results when collaborating with NLP experts. Continuing the metaphor, we now have the general knowledge so we can navigate the "fish market" and choose the best species—one that delivers high-quality nutrients and is ethically sourced, i.e. procured in a fashion that maintains the viability of the resource (language), rather than dynamiting the fishing grounds for spectacular but one-time hauls.

6.3 For computational linguists

First, we discovered a spectrum of skills among documentary linguists that supported their quick grasp of ML principles and ability to work with NLP models. For example, a prominent skill among descriptive linguists is complex pattern recognition, which is also a cornerstone of ML. The field of linguistics has traditionally placed less emphasis on statistical patterns of language usage and instead focuses on generalizing from specific patterns in order to describe a language's structure and from there building abstract theoretical models. Nonetheless, we find linguists readily embrace statistical methods and bring their expertise in pattern recognition and data analysis to bear once they see the value of a machine-in-the-loop approach for their goals.

Second, the pressure of academic publishing, or commercial interests for those in industry, may lead to the prioritization of novelty. At the same time what is novel for NLP may not be valued by another field. This disconnect in perceived common

[4] We are aware that the learning and outcomes would not require us to be physically in the same room, but personal interaction certainly helps in creating trust.

goals among academics may lead to miscommunication between computational linguists and documentary linguists who do not care about the novelty of models as long as they are relatively simple, work reliably, and reduce the annotation workload necessary to discover novel linguistic phenomena. We prioritized documentary linguists' immediate needs; even though models like CRF are not state-of-the-art, they were ideal for connecting principles of linguistics to ML concepts and better-equipped documentary linguists to continue using what they learned after the workshop.

The third lesson for computational linguists was not new to us, but bears repeating. An ethically operating NLP project using minority language data should entail a willingness to engage in long-term collaboration. Crucially, long-term collaboration allows one to assess the benefit not to only NLP research or documentary efforts but also to the language communities whose data we are using. How to elicit language data while giving value to the community has been discussed in linguistics literature for the past 50 years and more (D'Arcy and Bender, 2023). Collaboration with experienced documentary linguists is one way to discover how "fishing" for data might become a fair market.

6.4 For all participants

Finally, our key observation at the workshop was that genuinely listening to the divergent concerns of the computational and documentary linguists and adjusting one's approaches to accommodate each others' goals and felt needs was able to overcome prejudices based on prior less-than-optimal interactions. The paramount concern for the linguists was that the language data—collected together with the language communities—would not just disappear somewhere, to reappear as part of someone's research with no connection to or benefit for the language communities in question or in an application the communities would be expected to pay for. For the computational linguists, the primary concern was not to gain access to the data as such, but rather whether the documentary linguists would have sufficient resources to run the ML algorithms (e.g. access to GPUs), wherever the documentary linguists wished to keep the data, without needing the NLP experts' support and time to rerun and adjust the code. In this end, one positive experience where concerns were voiced and understood changed what both groups feel is possible for AI in documentary linguistics.

7 Discussion & Conclusion

The workshop proved to be successful at equipping linguists to do their own AI "fishing" for several reasons. First, the interaction allowed the linguists to close the existing gaps in their knowledge of ML and its application in endangered languages. Second, the workshop format allowed the linguists to put this knowledge into practice right away. They worked on solving real research problems both teams faced—the need for more morphological interlinearization and quick POS tagging. After the three-day workshop, participants had a trained model in their hands. Third, both teams had constant support from computational linguists, who helped to fix errors in data and gave valuable suggestions on model or workflow optimization.

The composition of the two research teams played a large role in the success of the model development. Each team had at least one linguist trained in Algonquian or Dene linguistics, and at least one with basic programming skills. This allowed both teams to 1) quickly identify and correct mistakes in the training data and the model; 2) devise and implement solutions tailored to each language; and 3) expand and incorporate new training data by correcting the models' predictions. Having NLP experts in the room reduced the time needed for troubleshooting and fostered confidence.

Although this workshop's main goal was to educate linguists, it was also an exciting and educative experience for the NLP experts. So many NLP tasks that documentary linguists face are as simple as "shooting fish in a barrel." Hence, in three days the NLP experts saw maximum positive impact. As a result, the impact of our AI workshop has gone beyond a three-day event, leading to further collaborations and grant applications.

By describing the results of our workshop, we want to emphasize that progress in NLP does not always depend on inventing new methods; rather, it often lies in the meaningful application of established methods to different languages. After all, each language brings new and often unique challenges to old tools. We hope that this workshop's outcomes will set a positive trend of impactful collaborations between documentary and computational linguists and lead to better communication between these two fields. The models might seem complicated and intimidating at first, but all participants discovered that linguists do not need a degree in computer science to use AI.

Acknowledgments

We are grateful to the Blackfoot and Dënë Sųłiné communities for the opportunity to work with their languages. In particular, we want to thank Niitsitapii communities in Alberta, Canada and the Clearwater River Dene Nation in Saskatchewan, Canada. This workshop was supported by SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages". Blackfoot language documentation work was supported by SSHRC IG 435-2021-0562 and SSHRC PDF 756-2022-0428. Dënë Sųłiné language documentation work was supported by SSHRC IG 435-2020-1197.

References

Eung-Do Cook. 2004. *A Grammar of Dëne Sųłiné (Chipewyan). Algonquian and Iroquoian Linguistics – Special Athabaskan Number, Memoir 17*. Algonquian and Iroquoian Linguistics, Winnipeg.

Alexandra D'Arcy and Emily M. Bender. 2023. Ethics in Linguistics. *Annual Review of Linguistics*, 9(Volume 9, 2023):49–69. Publisher: Annual Reviews.

ELAN (Version 6.7) [Computer software]. 2023. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. [link].

Rachel Ermineskin and Darin Howe. 2005. On Blackfoot syllabics and the law of finals. Paper presented at the 37th Algonquian Conference.

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

Donald G. Frantz. 2017. *Blackfoot grammar*. University of Toronto press.

Inge Genee. 2009. What's in a morpheme? Obviation morphology in Blackfoot. *Linguistics*, 47(4):913–944.

Inge Genee and Marie-Odile Junker. 2018. The Blackfoot Language Resources and Digital Dictionary project: Creating integrated web resources for language documentation and revitalization. *Language Documentation & Conservation*, 12:274–314. Publisher: University of Hawaii Press.

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

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Glenbow Museum. n.d. Kaitapiitsinikssiistsi / Traditional Stories.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 52–57, St. Julians, Malta. Association for Computational Linguistics.

Dominik Kadlec. 2023. A computational model of blackfoot noun and verb morphology. Master's thesis, Lethbridge, AB, Canada.

John Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning*, pages 282–289. Cited in Synthesis and paper for CRF.

Olga Lovick. 2020. *A Grammar of Upper Tanana, Volume 1: Phonology, Lexical Classes, Morphology*. University of Nebraska Press, Lincoln.

Sandra Áístainskiaakii Many Feathers, Brent Issapóíkoan Prairie Chicken, Wes Áínnootaa Crazy Bull, and David Osgarby. 2013. Aakíípisskani / the women's buffalo jump. In *Papers of the 48th International Conference on Salish and Neighbouring Languages*, volume UBCWPL35, pages 1–21. University of British Columbia.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-loop with documentary and descriptive linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Juri Opitz, Shira Wein, and Nathan Schneider. 2024. Natural language processing relies on linguistics.

Stephanie Carrol Rainie, Tahu Kukutai, Maggie Walter, Oscar Luis Figueroa-Rodriguez, Jennifer Walker, and Per Axelsson. 2019. Issues in open data - Indigenous data sovereignty. In T. Davies, S. Walker, M. Rubinstein, and F. Perini, editors, *The State of Open Data: Histories and Horizons*. African Minds and International Development Research Centre, Cape Town and Ottawa.

Lena Heavy Shields Russell and Inge Genee. 2014. *Ákaitsinikssiistsi: Blackfoot Stories of Old (First Nations Language Readers Blackfoot)*. University of Regina Press.

Statistics Canada. 2022. Mother tongue by geography, 2021 [data visualization tool].

Sowmya Vajjala. 2021. Teaching NLP outside linguistics and computer science classrooms: Some challenges and some opportunities. In *Proceedings of the Fifth Workshop on Teaching NLP*, pages 149–159, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jan Philip Wahle, Terry Ruas, Mohamed Abdalla, Bela Gipp, and Saif Mohammad. 2023. We are Who We Cite: Bridges of Influence Between Natural Language Processing and Other Academic Fields. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12896–12913, Singapore. Association for Computational Linguistics.