Lemmatization of the Corpus of Historical Mapudungun

Nikita Moghe, Rimvydas Rubavicius, Dan Wells, Irene Winther
Centre for Doctoral Training in Natural Language Processing
School of Informatics, University of Edinburgh, United Kingdom
{nikita.moghe,rimvydas.rubavicius,dan.wells,irene.winther}@ed.ac.uk

Abstract

We present a framework for the lemmatization of a morphologically-rich language in a low-resource setting, working with a small number of annotated texts from a corpus of historical Mapudungun. The goal is to extend the annotation of lemmas to the entire corpus for the purpose of language documentation. In this framework, we apply text normalization to address orthographical variation in the corpus and use three separate models based on recent work to predict output lemmas. The framework achieves 77.2% exact match accuracy on the test set which is the most representative of the corpus. This automated prediction of lemmas can help speeding up manual annotation of the remaining documents in the corpus. This work demonstrates one way in which practices in Natural Language Processing can be used to aid language documentation efforts.

1 Introduction

The current paper explores the use of NLP tools in the context of language documentation, which involves making linguistic material accessible for research, teaching and cultural interest. This typically requires linguistic annotation of a corpus of linguistic materials, which is a time-consuming and expensive process, as input from linguists knowledgeable in the target language is needed. While automatic methods exist for linguistic annotation, such as part-of-speech tagging, these often require large amounts of already-tagged data to train models.

In this work, we present a framework for aiding the annotation effort for a new corpus of historical Mapudungun – a morphologically-rich language of South America – with very little labelled data. The Corpus of Historical Mapudungun (CHM) aims to make linguistically-annotated Mapudungun texts accessible to facilitate diachronic studies of language change, for example by making a given word searchable across multiple texts using an information retrieval system. A key component for building such a system, especially for morphologically complex languages like Mapudungun, involves full lemmatization of the corpus. Lemmatization is the process of providing a canonical representation (dictionary lookup forms) for each surface form in a document.

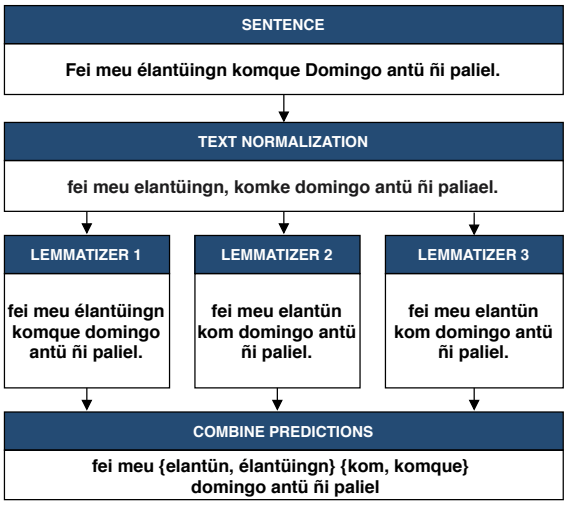


Figure 1: Outline of proposed lemmatization framework including multiple components.

Several challenges arise in the lemmatization of the CHM. First, as the texts are compiled in a 300-year period with multiple authors, the corpus data shows variation resulting from historical processes of linguistic change, dialectal variation, as well as differing transcription systems. Second, the complex verbal morphology results in a large number of possible surface forms for any given word, even if orthogra-

phy or time period were standardized. Third, the amount of annotated training data is limited, with around 4,000 word tokens currently tagged.

We combine a text normalization system to deal with data variability with three separate lemmatization models, as outlined in Figure 1. Predictions from the individual models can be ranked and presented to a linguistic expert for review and selection of the appropriate lemma in each case.

In what follows, we first describe the language and its morphological features in Section 2, with the challenges these pose for lemmatization. We then describe the corpus and how we address the issue of data variation in Section 3 before we describe the three models used for lemmatization in Section 4. We discuss our results which motivate the use of three models in Section 5, and consider future extensions to the annotation framework in Section 7.

2 Language Description

Mapudungun is an indigenous language spoken in Chile and Argentina. Although there are approximately 200,000 speakers of Mapudungun, it is considered endangered due to poor transmission between generations of speakers (Molineaux, 2019). The language is also presumed to be an isolate and thus does not have any known linguistic relatives. This means that, unlike some other low-resource settings, we cannot use data from related languages as a starting point for our models.

Mapudungun is heavily agglutinative, meaning that it has a complex morphology where each morpheme, such as a suffix, represents a single grammatical function. These morphemes are then "glued" onto a stem to form more complex expressions, as shown in example (1).

(1) *kude -ke-fu-i-ng-u*
play/bet-HABIT^{14}-PAST^{8}-IND^{4}-3^{3}-DUAL^{2}
'the two of them used to bet'

The most distinct morphological complexity can be found in the verbs: there are about 100 suffixes that can be added to a verb stem to convey grammatical information such as valence, aspect, modality, negation, person, number and tense. A Mapudungun verb has at least one and usually no more than ten suffixes attached to what Smeets (2008) distinguishes as 36 more or less fixed "suffix slots" that follow the verb root, an example of which can be seen in example (1) above, where the superscripts indicate the slot position of each suffix in this example. There are no irregular verbs, but they may be complex, containing more than one verb root (e.g. *anü-püra* 'to sit up'-sit.down-go.up-), in which case determining the correct lemma can be difficult. The other parts of speech have a relatively simple morphology, with the exception of nominals, which may have some suffixation, compounding and reduplication. The complex verbal morphology also allows for a relatively free word order.

From an NLP perspective, a language like Mapudungun will have significantly more lexical types (unique words) in its vocabulary compared to an analytic language with little inflection, such as English. Thus, the number of unique surface forms of a lemma is larger, at least for verbs.

Considering the relatively fixed pattern of the verbal suffixes, a rule-based approach could be considered for lemmatizing Mapudungun verbs. Here, a surface form is parsed from right to left (from the outermost suffix slot towards the root), identifying each suffix to find the root which could be a potential candidate for the correct lemma. As an example, in order to find the lemma *kuden* 'play/bet 'of the verb in example (1) above, the five suffixes -ke, -fu, -i, -ng, and -u need to be identified. However, there are several challenges for this approach:

1. Some slots may have zero fillers (-∅), i.e. suffixes that are not overtly marked.

2. A slot may have different, mutually exclusive suffixes, e.g. slot 2 can have either -∅, -i, -u, -iñ or -ün, indicating number.

3. The suffixes are not fixed in length due to morphophonological constraints, for instance a suffix -n could also be -ün depending on the preceding sound.

4. The root is not equal to the lemma for most verbs, as exemplified in (1) where the root is *kude* and the lemma *kuden*.

5. A separate set of rules is needed for the other parts of speech.

As a consequence, other approaches to lemmatization need to be considered.

3 Corpus Description

The Corpus of Historical Mapudungun[1] is an ongoing collection of Mapudungun texts spanning the period 1606–1930. The Mapudungun data represented were compiled across that time period by multiple authors with different focuses, for example missionaries, explorers, military men and anthropologists, and therefore cover multiple domains such as descriptive grammars, vocabularies and phrasebooks, sermons and Christian doctrinal texts and records of dialogues, stories and songs from native speakers (Molineaux, 2019). Digitized materials and archival images for these sources held by the *Biblioteca Nacional de Chile* and the *Archivo Rodolfo Lenz* in Santiago, Chile were passed through an optical character recognition (OCR) system powered by Google Cloud Vision using the Digital Humanities Dashboard (Tarpley, 2017). This mixture of time periods, authors and domains, as well as errors in OCR output, present multiple sources of variation in the data which must be accounted for.

3.1 Difficulties for lemmatization

Given the methods and sources used to compile the CHM, the surface form of a particular lemma for a particular set of morphological features could potentially vary across different documents for three reasons:

1. Historical change or dialectal differences in the phonetic form of inflectional markers reflected in the spelling
2. Differences in orthographic systems used by different authors
3. OCR errors producing incorrect output character sequences

In each case, we would like to retrieve the same underlying lemma for multiple different spellings of a given inflected form, regardless of the source of the spelling mismatch between examples.

We may note that written differences reflecting historical phonological change is precisely the kind of variation the CHM hopes to make accessible for study. Differences of this kind would result in different surface forms in texts from different time periods even if the characters used to represent particular phonemes were completely consistent between the two. Author-specific orthographic systems, on the other hand, can result in the same sound being represented with different character sequences in different texts. The choice of orthographic system is generally consistent across multiple texts from a single author, however, and the differences between authors' chosen transcription systems can be accounted for systematically. Mistakes in OCR outputs are occasionally consistent (e.g. the 'long s' <ſ>consistently output as <f>) but can also result from visual noise such as marks or visible text from the other side of the digitized page.

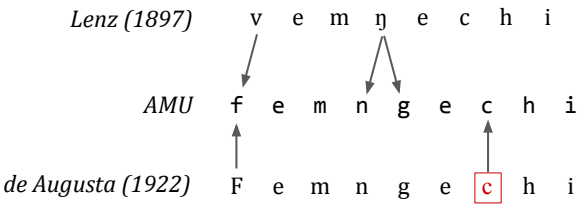


Figure 2: Text normalization from OCR characters to AMU orthography.

There is no universally-accepted standard orthography for modern Mapudungun, with three spelling systems in use today (Ager, 1998). Of these, the *Alfabeto Mapuche Unificado* (AMU) is generally used in linguistic work on Mapudungun, and this is the system used for gold-standard lemmas in the annotated portion of the CHM. Each author of the historical Mapudungun data in the CHM also uses their own specific orthographies, with varying degrees of divergence from the AMU standard. Rather than place the burden of learning mappings from each orthographic system into AMU on our models in addition to the lemmatization task, we take a preprocessing approach which converts input text to AMU orthography before passing to the model, both during training and at test time.

Figure 2 shows two instances of the same underlying word *femngechi* drawn from two CHM documents with different authors. This example shows different orthographic choices of the two authors which may also differ from the

[1] https://benmolineaux.github.io/. Last accessed: 2020-01-08

Document set	Texts	Authors	Tokens
Full CHM	36	21	400k
Phrase-level XML	9	6	40k
Fully-annotated	2	2	4k

Table 1: Outline of CHM documents. Each row represents a subset of the one above.

Data set	Tokens	Types	Lemmas
Train	3,226	1,209	731
Validation	363	240	199
Standard Split	377	242	207
First Sentences	727	391	283
Unseen Words	283	268	214

Table 2: Data sets used.

modern AMU, as where Lenz uses the single character <ŋ>to represent velar nasals rather than the AMU digraph <ng>. We also see the possibility of OCR errors such as the Cyrillic character highlighted in red which has been substituted by the OCR model in place of the expected and near-identical Latin character <c>. Any such errors which result in characters outside the expected set of alphabetic and punctuation characters appearing in document transcriptions are corrected as part of the text normalization step. We also remove capitalization throughout, as at the beginning of the token drawn from de Augusta.

3.2 Data sets and annotation

As work progresses on the creation of the CHM, its constituent documents are in different stages of preparation. Table 1 shows the current state of the corpus, with each row representing a subset of the documents in the row above it.

Fully-annotated documents have already been processed by a linguistic expert, and are provided in XML format following the Text Encoding Initiative guidelines and schema (TEI Consortium, 2007). As part of the linguistic analysis, the body of each text has been tokenized into word-level elements with attributes including the corresponding lemma, part of speech (POS) and English translation. Word elements have also been decomposed into their constituent morpheme elements, with attributes including the base form of the morpheme (useful in case of morphophonlogical processes which may the surface form) and its grammatical function.

These fully-annotated documents provide the wordform-lemma pairs required for model training and validation. The two texts were combined and sentences randomly sampled to create an 80% training split and 10% validation split. The remaining 10% was held out as a test set, which we refer to below as *Standard Split*.

Considering the final application of our system, namely extending the annotation of CHM documents, we would ideally have a wider sample to evaluate on besides sentences drawn from the two training documents, which only contain data from a limited time period and two different orthographic systems. We therefore requested additional data annotation from the compiler of the CHM, covering the first few sentences (between 2–8 depending on length) of each remaining document in the phrase-level XML set. The resulting annotated sentences are referred to as the *First Sentences* test set below, and provide wider coverage over the orthographic systems and historical periods represented in the full CHM.

Finally, we have a test set of *Unseen Words* randomly sampled from the phrase-level XML documents, and which are not seen in either of the fully-annotated training documents. Between the *First Sentences* and *Unseen Words* test sets, we are able to test the generalization performance of our lemmatization systems beyond the specific transcription conventions and historical period represented in the training data. Table 2 summarises the data size and composition of each data set, after applying text normalization.

4 Methodology

Lemmatization can be formulated as a supervised learning problem in which, given a data set $\mathcal{D} = \{(\mathbf{x}^{(n)}, \mathbf{y}^{(n)})\}_{n=1}^{N}$ of wordform-lemma pairs, a function f can be learned to map surface forms to the corresponding lemmas. Each wordform and lemma consists of a sequence of characters: $\mathbf{x}^{(n)} = x_1^{(n)}, x_2^{(n)}, ...x_I^{(n)}$ $\mathbf{y}^{(n)} = y_1^{(n)}, y_2^{(n)}, ...y_J^{(n)}$ of length I and J, respectively.

It has been shown that considering contextual information may help to disambiguate in cases where a single wordform may have multiple lemmas (Bergmanis and Goldwater, 2018), yet it was decided not to follow such an approach to lemmatize CHM due to low amount

of training data and a relatively free order of the words in the language. As such, all models perform lemmatization of single wordform inputs at a time.

4.1 Text preprocessing

We implement text normalization using the OpenGrm Thrax grammar development tools (Roark et al., 2012) to produce a set of weighted finite-state transducers mapping from each of the CHM author-specific orthographic systems into AMU. As described in Section 3, these grammars also lowercase all characters in the input text and correct transcription errors from the OCR process.

While the remaining unannotated documents in the corpus are provided in a similar XML format as the training data, they are only structured down to the phrase or sentence level. When processing these documents, we perform basic tokenization on whitespace and punctuation to split these lines into the word-level tokens as expected by our lemmatization model, taking into account the use of hyphens to indicate compounding and apostrophes as part of certain graphemes in several Mapudungun orthographies. For annotated test documents, we take tokens as given in the word-level XML structure based on human linguistic analysis.

4.2 Baselines

Copy

This baseline uses an identity function which predicts lemma being the same as the wordform by copying it.

$$f(\mathbf{x}^{(n)}) = \mathbf{x}^{(n)} \tag{1}$$

This is a sensible baseline as it captures wordforms that match their lemmas. This is typical for pronouns, prepositions, and other closed class words. Yet, it cannot capture morphological variations of words sharing the same lemma.

Most frequent

This baseline uses memorization of the training data by creating a look-up table T from wordforms to their lemmas. In case of the ambiguous wordform-lemma mapping, the most frequently occurring lemma for a given wordform is stored, breaking ties randomly. If the wordform has not been observed in training data, we fall back to the copy baseline.

$$f(\mathbf{x}^{(n)}) = \begin{cases} T[\mathbf{x}^{(n)}] & \text{if } \mathbf{x}^{(n)} \in \mathcal{D} \\ f(\mathbf{x}^{(n)}) & \text{otherwise} \end{cases} \tag{2}$$

This baseline correctly lemmatizes wordforms seen in training, tackling some of the morphological complexity in the data, but cannot generalize well to unseen wordforms even though they would exhibit similar morphological patterns as seen in the training data.

4.3 Models

LEMMING

Lemming (Müller et al., 2015) creates a list of possible lemmatization rule templates $\mathbf{r}^{(n)}$ for each wordform by considering edit trees between wordforms and lemmas, as illustrated in Figure 3a. Using these rule templates, a discriminative classifier is learned to choose which rule to apply for a particular wordform at test time:

$$f(\mathbf{x}^{(n)}) = \arg\max_{\mathbf{r}^{(n)}} P(\mathbf{r}^{(n)}|\mathbf{x}^{(n)}) \tag{3}$$

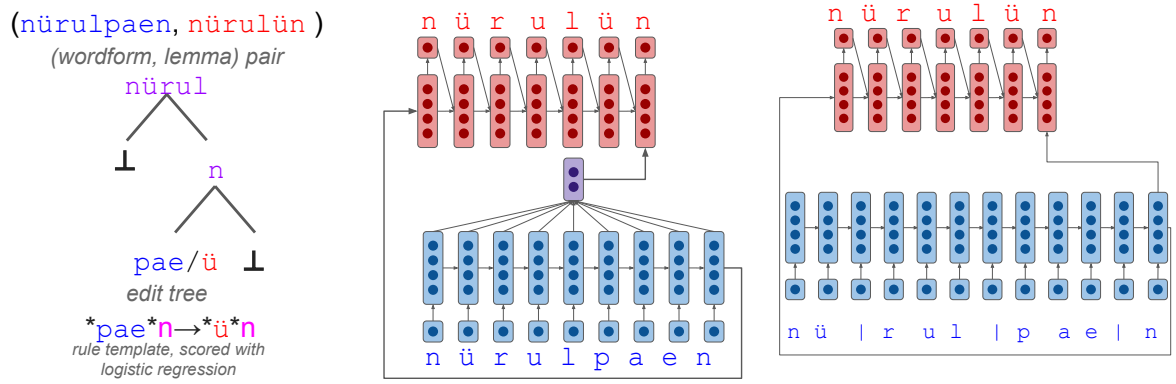
Lemming also learns a discriminative classifier to perform part of speech (POS) tagging. While we do not use these predicted tags ourselves, they are a useful additional output for the final annotated corpus.

The main disadvantage of Lemming is that it only applies one rule at a time while there may be multiple morphological phenomena occurring at once. This can be partially solved by either increasing training data or by using a more flexible model.

Lematus

Lematus (Bergmanis and Goldwater, 2018) is a neural model that uses recurrent sequence-to-sequence model with attention, commonly used in neural machine translation (Bahdanau et al., 2015). The sequence of wordform characters is treated as a source language sentence and the sequence of lemma characters as target language sentence, as shown in Figure 3b.

Lematus is flexible enough to be able to handle multiple morphological phenomena occurring at once.



(a) Lemming (b) Lematus (c) HMA

Figure 3: Lemmatization models used. Input is marked in blue and the output is marked in red.

Hard Monotonic Attention

The hard monotonic attention (HMA) model of Aharoni and Goldberg (2017) was originally used for the task of morphological re-inflection, generating inflected wordforms given a lemma. This model differs from Lematus in the attention mechanism used, with hard attention considering only a single character at a time rather than a weighted combination of all characters in the input as in Bahdanau et al. (2015). This hard attention mechanism is constrained to proceed monotonically through the input string, reflecting the serial nature of morphological suffixation in an agglutinative language such as Mapudungun. We reuse this architecture for the task of lemmatization by inverting inputs and outputs, thereby learning to predict lemmas given wordforms, as shown in 3c.

Additionally, it was found useful to augment the input sequence by including the morpheme boundaries (denoted by | symbol) for each wordform. Morpheme boundaries are included as part of the annotated the training data, whileand are used to learn a separate HMA model to map wordforms to their segmented versions at test time. Note that similar segmentation was also explored with Lematus but proved to be unsuccessful.

HMA is both flexible in handling various morphological phenomena and able to effectively use additional information provided in the corpus.

4.4 Combining Predictions

For the final part of the lemmatization framework, lemma predictions are ranked to bootstrap the annotation process. For each wordform, we aggregate lemma predictions, leading to a hypothesis list $[h_1, h_2, h_3]$. Elements in this list are ranked based on these rules:

- Neural model predictions are ranked based on the heuristic of their negative log-likelihood score: the higher score model will be h_1 and the lower score will be h_2.
- As model scores are not accessible in the implementation of Lemming used, it is ranked as the last hypothesis h_3. This also reflects the expectation that these predictions are likely to be less accurate than the neural models due to the reduced flexibility of Lemming.

Given duplicates in the hypothesis list, we keep the hypothesis with the highest rank.

4.5 Evaluation metrics

We evaluate our lemmatizers on two metrics: Exact Match accuracy (EM) on the entire test sets and average Minimum Edit Distance (MED) over the faulty predictions.
Exact match accuracy measures the percentage of correctly predicted lemmas. It provides an estimate of the percentage of the corpus that can be automatically lemmatized.
To check if the incorrect predictions can still be useful to the corpus lemmatization process, we report the average MED of the erroneous predictions for all the experiments. MED provides the minimum number of edits that an annotator needs to make from the faulty prediction to the ground truth. These edits include the

Text	Raw Types	Norm. Types
Lenz (1897)	1,247	1,128
de Augusta (1922)	410	386

Table 3: Data counts in raw and normalized annotated documents.

Test scenario	Raw EM	Norm. EM
Cross-document	55.9	73.2
Standard Split	69.8	74.5

Table 4: Lemma exact match percentages for raw and normalized test sets.

insertion, deletion or replacement of individual characters.

5 Results and Discussion

In this section, we first report the benefits of using text normalization followed by a detailed discussion on the empirical performance of different lemmatization models.

5.1 Impact of text normalization

To gauge the usefulness of applying text normalization to the training texts, we initially evaluated the Most Frequent baseline by training on the larger of the two fully-annotated documents and testing on the other, before and after normalizing text in each document. Table 3 shows data counts for this cross-document scenario.

Lemma exact match results for this scenario are shown in Table 4, along with pre- and post-normalization results using the mixed-document training and *Standard Split* test set. We saw a large improvement in the baseline results for the cross-document scenario, as particular wordforms spelled one way by the author of the training document (Lenz) but differently in the test document (by de Augusta) are united in the AMU orthography targeted by the text normalization process. After normalization, the number of overlapping word types between the two documents increases from 59 to 83, and that increased overlap directly allows wordform-lemma statistics calculated on one document to be referenced when lemmatizing the other using this baseline function. We also see a small benefit on the *Standard Split* when compared to raw wordforms.

This analysis provided an early result of the possible benefits of text normalization, motivating us to continue with this approach while constructing the rest of the framework described in this paper. The development of the text normalization system will also benefit ongoing efforts of cleaning up the text documents in the CHM, after identifying numerous OCR errors as described in section 3.2.

5.2 Quantitative Evaluation

After verifying the importance of text normalization, all the lemmatizers were trained and evaluated on their normalized versions. The performance of all the models on different test sets is described in Table 5.

As seen in the table, the baselines *Copy* and *Most Frequent* perform worst overall, suggesting that using such simple heuristics is not sufficient to build a lemmatizer for this task. Lemming improves over the baselines suggesting that edit trees are helpful in capturing morphological differences between wordforms and lemmas to a certain extent. Neural models have the best performance, with *Lematus* giving the best EM performance and *HMA* the best MED. This is in line with the empirical observations in Bergmanis and Goldwater (2018), where Lematus' performance was competitive or better than Lemming across several languages.

From Table 5, we observe that all the models perform the best on the *Standard Split*, which is expected as the words from the test set were taken from the same source documents. The models exhibit a lower performance on the generalization test set *Unseen Words*. We hypothesize this behaviour as a result of Mapudungun's complex verbal morpohology, where verbs make up 42.4% of items in this test set compared to 26.3% in the *Standard Split*, as well as the generally harder task of making unseen predictions.

It is interesting to note that the edit distance has a similar performance across all the test sets for the neural models. This indicates that the erroneous predictions from these models can still be useful to the annotators as an average 2-3 edits at the character level are required to obtain the correct lemma. Moreover, the exact match of the models on the *First Sentences*, the test set which is most representative of our final application, is adequate enough to provide a set of initial predictions for the remaining untagged corpus.

Method \Dataset	Standard Split		Unseen Words		First Sentences	
	EM ↑	MED ↓	EM ↑	MED↓	EM ↑	MED ↓
Copy	44.6	2.7	25.7	3.3	47.3	2.4
Most Frequent	74.5	3.9	30.6	3.4	62.9	2.9
LEMMING	80.9	3.3	33.9	3.0	65.8	2.4
Lematus	**86.5**	2.4	**45.2**	2.6	**69.7**	2.5
HMA	83.0	**2.2**	44.2	**2.5**	69.2	**2.4**
Combined	**91.2**	**1.8**	**59.7**	**2.0**	**77.2**	**1.8**

Table 5: Evaluation of all the methods on the three test sets. Combined at the end of the table combines the hypotheses from LEMMING, Lematus, and HMA.

Finally, we evaluate the performance of our framework as a whole by considering the benefits of presenting multiple hypotheses to a linguistic annotator. We evaluate the predictions from all the three models together against the gold standard such that the predicted lemma is considered correct if at least one model predicts the gold lemma. The reported MED for examples for which no model is able to predict the correct lemmas is calculated by taking the lowest MED from the three incorrect predictions and computing the average. Thus, the lower bound for the combined predictions is equivalent to the EM of *Lematus* and MED of *HMA* which are the best performing models on the respective metrics.

While this demonstrates that we could potentially show the annotator the correct lemma in a majority of cases, there is an open question whether or how much having multiple hypotheses to choose from would actually speed up the annotation process. One method we use to approach this is by ranking hypotheses as discussed in Section 4.4, but we leave evaluation of annotation speed improvements for future work.

5.3 Error Analysis

For performing error analysis, we first categorize our results in terms of POS-tags. The results are reported on nouns and verbs, and the other parts of speech are categorized as "Others", as nouns (24.2%) and verbs (26.6%) are the most frequent and exhibit the most variation in the corpus. Typically, for all the models across all the test sets, the performance on nouns was higher than the verbs, with an exception of HMA on *Unseen Words*. This indicates that verbs in Mapudungun indeed have a complex morphology as discussed in Section 2. The "Others" set exhibited the highest performance across all models and all test sets, except Lematus on *Unseen Words*. This was expected as most of the closed class words were present in the training set and the open class words in this category, such as adjectives, do not display significant variation. For the sake of brevity, we report the exact match accuracy on *First Sentences* in Table 6 and the rest in the Appendix.

Model	Nouns	Verbs	Others
LEMMING	66.7	30.2	78.5
Lematus	61.8	55.4	79.2
HMA	62.8	51.1	79.3

Table 6: Performance of different models across nouns, verbs and other parts of speech on the *First Sentences* test set.

We also inspected our results manually to identify if the models fail on a particular set of words or have frequent error patterns. As exemplified in Table 7, the models produce erroneous predictions on different words and therefore one model could still yield a correct prediction if the other models fail. This explains the increase in performance across the three test sets (4.7%, 14.5%, 7.5%) by using combined predictions (see Table 5).

wordform \predicton	LEMMING	Lematus	HMA
pepiavui	**pepin**	pen	pafün
duŋu-librolimi	düngu-libroln	**düngulün**	düngulin
küpalelərkeneu	küpalelərkn	küpalün	**küpan**

Table 7: Sample predictions from the three models for three examples. The words in **blue** are the correct lemmas for the given surface forms respectively.

6 Related Work

Lemmatization is a widely studied task in NLP as it is an important component for downstream applications like machine translation (Fraser et al., 2012), parsing (Björkelund et al., 2010), and keyphrase extraction (El-Shishtawy and Al-sammak, 2012) to name a few. The approaches to solve lemmatization involve rule-based morpohological analysis as well as borrowing techniques from machine learning. Chrupala (2006); Müller et al. (2015) treat lemmatization as a classification problem where the task is to find the correct edit tree that should be applied to convert the surface form to its corresponding lemma. Lemmatization also uses techniques from its reverse task of morphological re-inflection where the target is to predict a surface form given its lemma and morphological properties. Faruqui et al. (2016); Kann and Schütze (2016) posed the task of morphological re-inflection as a neural sequence-to-sequence problem over learning a sequence of characters. Lematus and HMA are extensions of the work in Kann and Schütze (2016). Comparing results on recent low-resource lemmatization challenges, we note performances around 60–65% for the top 5 participants of the low-resource lemmatization portion of the CoNNL 2018 Shared Task (Zeman et al., 2018).

Littell et al. (2018) present a number of NLP applications applied to documentation and revitalization efforts for indigenous languages of Canada. Working with these languages poses similar challenges to Mapudungun, with many exhibiting complex morphology and having limited training data available. Micher (2017) parallels our work for Inuktitut, which shows very similar morphological behavior to Mapudungun as well as having non-standard orthography. That work is more successful, however, given an existing rule-based morphological analyzer for Inuktitut providing upwards of 150k word types and annotations then used to train a more robust recurrent neural network-based analyzer.

7 Conclusion and Future Work

This work provides the initial steps to aid the annotation efforts of the Corpus of Historical Mapudungun for the purpose of language documentation. We have presented a lemmatization framework to address the challenges encountered in this corpus. We find that text normalization is crucial when parsing texts across different time periods and authors to address some of the variation in the data. We validate the usefulness of three existing lemmatization techniques for Mapudungun in the low resource setting. As this work is the first step towards fully annotating the corpus, there are interesting directions to improve upon this work:

- **Evaluation**: Qualitative evidence is required in order to evaluate the extent to which the annotation process is enhanced by the proposed framework. The presentation of the combined predictions may be modified by performing user studies and consulting annotators to better match the expected functionality. Online learning (Poggio and Rosasco, 2015) could be considered to further improve performance based on the feedback an annotator may provide on the combined predictions. Moreover, as each hypothesis list has a score, a threshold could be found to allow some of the hypotheses to be taken as a ground truth and thus not requiring selection by an annotator.

- **Further annotations**: In the intended IR system, the search can be specified by a lemma, as well as part-of-speech, morphemes and their types, spelling, and the English correspondence. This work lays the foundations for part-of-speech tagging as LEMMING provides predictions for them, as well as for morpheme detection as this corresponds to the segmentation performed by HMA.

- **Unsupervised learning**: Our framework only considered supervised learning methods while we can still explore unsupervised and semi-supervised learning methods. For instance, one could learn a language model over lemmas to score predictions made by all three lemmatization models.

Acknowledgments

We would like to thank Sharon Goldwater for advising us along the way, Benjamin Molineaux

for providing insights and additional annotations of the Corpus of Historical Mapudungun, as well as members of the CDT in NLP and the organizing team for joining us in useful discussions and providing feedback during the course of this project.

Contributions

Responsibilities of each member of the team:

- Nikita: HMA experiments and implementing rule-based approach (discussed in Section 2) with Irene. Writing up section 5, some parts in 6 and 4.
- Rimvydas: Lematus experiments, writing up sections 4 and 7, generating visualizations.
- Dan: LEMMING experiments, text normalization and parsing XML data. Writing up sections 1, 3 and 6 and proofreading throughout.
- Irene: linguistic features of Mapudungun, error analysis, identifying rules for rule-based approach. Writing up sections 1, 2, and parts of Section 7.

References

Simon Ager. 1998. Mapuche. In *Omniglot - writing systems and languages of the world*. https://www.omniglot.com/writing/mapuche.htm. Last accessed: 2020-01-08.

Roee Aharoni and Yoav Goldberg. 2017. Morphological Inflection Generation with Hard Monotonic Attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2004–2015, Vancouver, Canada. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Toms Bergmanis and Sharon Goldwater. 2018. Context Sensitive Neural Lemmatization with Lematus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1391–1400, New Orleans, Louisiana. Association for Computational Linguistics.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *COLING 2010, 23rd International Conference on Computational Linguistics, Demonstrations Volume, 23-27 August 2010, Beijing, China*, pages 33–36.

Grzegorz Chrupala. 2006. Simple data-driven context-sensitive lemmatization. *Procesamiento del Lenguaje Natural*, 37.

Tarek El-Shishtawy and Abdulwahab Al-sammak. 2012. Arabic keyphrase extraction using linguistic knowledge and machine learning techniques. *CoRR*, abs/1203.4605.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 634–643. The Association for Computational Linguistics.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling inflection and word-formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 664–674, Avignon, France. Association for Computational Linguistics.

K Kann and H Schütze. 2016. MED: The LMU system for the SIGMORPHON 2016 shared task on morphological reinflection. *Proceedings of the 14th SIGMORPHON Workshop*.

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

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Benjamin Molineaux. 2019. The Corpus of Historical Mapudungun: Digital Tools for New-World Language Change. http://www.homepages.ed.ac.uk/bmolinea/talk/dh2019/. Last accessed: 2020-01-08.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint Lemmatization and Morphological Tagging with Lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

Tomaso Poggio and Lorenzo Rosasco. 2015. Machine learning: a reguliarization approach.

Brian Roark, Richard Sproat, Cyril Allauzen, Michael Riley, Jeffrey Sorensen, and Terry Tai. 2012. The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, pages 61–66, Jeju Island, Korea. Association for Computational Linguistics.

Ineke Smeets. 2008. *A Grammar of Mapuche*. Number 41 in Mouton Grammar Library. Mouton de Gruyter, Berlin.

Bryan Tarpley. 2017. Breakdowns in Machine Reading: Attempting to De-privilege Modern English Print with the Power of Supercomputing and the DH Dashboard - UNT Digital Library. https://digital.library.unt.edu/ark:/67531/metadc1010762/. Last accessed: 2020-01-08.

TEI Consortium. 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange. https://tei-c.org/Vault/P5/1.0.0/doc/tei-p5-doc/en/html/. Last accessed: 2020-01-10.

Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics. Low-resource lemmatization results retrieved 2020-01-08: https://universaldependencies.org/conll18/results-lemmas.html#low-resource-languages-only.

A Extended Results

Model	Nouns	Verbs	Others
Lemming	53.6	17.5	55.6
Lemmatus	52.4	38.3	48.1
HMA	40.2	43.3	49.4

Table 8: Performance of different models across nouns, verbs and other parts of speech on the *Unseen Words* test set.

Model	Nouns	Verbs	Others
Lemming	85.1	48.5	96.2
Lematus	85.1	69.7	96.2
HMA	79.8	66.7	93.5

Table 9: Performance of different models across nouns, verbs and other parts of speech on the *Standard Split* test set.