Cognate Prediction for Under Resourced Languages
Information Technologies for Marginalised Languages

Evan Chapple
University of Lorraine
evanpkchapple@gmail.com

Amelie Knecht
University of Lorraine
aknecht@kabelmail.de

Ezgi Basar
University of Lorraine
ezgibasar42@gmail.com

Rachel Atherly
University of Lorraine
rachelatherly98@gmail.com

Mehsen Azizi
University of Lorraine
Mehsen-azizi@live.com

Mina Oulhen
University of Lorraine
m.oulhen2@gmail.com

Tiankai Luo
University of Lorraine
clarencetiankailuo@gmail.com

Meryem Bouziani
University of Lorraine
meryembouziani729@gmail.com

Abstract

This project aims to predict cognates for marginalised languages. A deep learning model has been trained on a Spanish-French cognate corpus transcribed into the International Phonetic Alphabet (IPA) to predict the correct Spanish cognate of a given French word. The French-Spanish cognate pairs were used as a proof of concept of the model to then be applied to Finnish and Inari Saami, a severely endangered language. An attention mechanism is used to improve results and to increase the explainability of the model. In reconstructing missing vocabulary in low-resource languages, this study thereby lays essential groundwork for revitalizing endangered languages.

1 Introduction

As of 2022, 50% of the world's languages became endangered. Each year, even more are lost, despite preservation efforts from linguists and native speakers alike, and many more will fall into complete extinction over the next decades. Many critically endangered and dead languages lack strong written records which leads to vocabulary loss and holes in the lexicon, making it difficult to revive the language in its entirety. For this reason, there is a constant need for research into new methods for language reconstruction. Tracking phonetic changes between cognates in closely related languages is one way to try and fill in these missing gaps. Cognates are vocabulary items of two related languages that can be traced back to a common ancestor. While following patterns of regular phonetic shifts, they can remain phonologically close. By identifying cognates in two related languages, these changes can be analyzed and then utilized to make predictions on how other words may follow the same pattern(s). These predictions can then be used to attempt to reconstruct vocabulary that has been lost to time. This study aims to build a framework for predicting missing vocabulary in endangered languages. To this end, as a proof of concept, a neural network model has been trained to predict a cognate in French given the word in Spanish, both of which are two closely related, roman languages. Yielding promising results in this first step, the model was then applied on the target language Inari Saami and its close relative Finnish. Inari Saami is a severely endangered language from the Uralic language family with only a few hundred speakers left. The lack of major revitalization campaigns leaves Inari Saami with significant gaps in the vocabulary. Our model could provide suggestions on how to fill these gaps and thereby contribute to future efforts to revitalize and maintain the language, culture and people. The paper is structured as follows: A more detailed description of cognate identification in historical linguistics and an overview of related work in the area in 2 and 3; The French-Spanish corpus and Inari-Finnish corpus used and the different challenges faced preprocessing the language data in 4; The model description and training process in 5; The results it yielded and their discussion in 6 and 7; The carbon footprint of training the model in 8 and plans for future work in 9.

2 Cognate Identification

Cognates are words from two different languages that descend from a common ancestor. The identification of cognates is an important task in historical linguistics. Historical linguists often make use of the comparative method to identify cognates and determine how languages are related. In this method, linguists examine words and languages comparatively to infer the hidden phonological and morphological processes that govern their linguistic lineage (Hall and Klein, 2010). Cognates can sometimes be traced back to a proto-word in the ancestral language by extrapolating via systematic sound changes observed over a number of languages (Meloni et al., 2021). This sound correspondence refers to how a phoneme in one language can relate to a phoneme in another related language. Cognate words exhibit sound correspondences because they are the result of a series of regular phonological changes which since the divergence from the common ancestor language (Agarwal and Adams, 2007). The French sound corresponding to the Latin /k/ is reliably predictable based on the environment, for example; /k/ remains as /k/ before [o] and in clusters, turns into /s/ before front vowels, and turns into /ʃ/ before [a]. Over the course of the diversification, different phonological rules develop for each language. In the case of Spanish, the Latin /k/ would turn into /θ/ before front vowels and remain as /k/ before [a]. (Anttila, 1989) Since sound changes are often regular, in a given cognate set, the Spanish /θ/ is expected to correspond to the French /s/ before front vowels. This makes it possible to predict cognates in language pairs using the comparative method. The manual approach to cognate identification relies on the historically informed judgments of trained linguists. Automatic cognate identification tasks also tend to use paired word lists and operate on word similarity; however, it is usually the case that automatic approaches only use surface resemblances without directly accounting for sound correspondences due to regular sound change. Therefore, automated approaches have the benefit of making quick and far-reaching judgments often at the cost of the accuracy offered by manual cognate identification.

3 Related Work

Automatic cognate identification and alignment has been the subject of numerous studies. However, few have worked with marginalised languages which are often under resourced which leads to little etymological knowledge and/or data. Linguists have historically used phonetic, orthographic and/or semantic approaches to identify and connect cognates. Different approaches have been developed to maximise preciseness of varied models. Below is an overview of some papers that treated related topics and have influenced the work done in this paper. (McCoy and Frank, 2018) tested three different methods for weighting edit distance algorithms, a method commonly used to relate cognates: feature-based edit distance, char2ev edit distance, and cognate-based edit distance. They compared those methods against the Levenshtein edit distance and showed that cognate-based edit distance performed the best for information transfer from a high-resource language to one of its low-resource relatives. (Fourrier et al., 2021) modeled the task of predicting the form of a cognate in another language as a low-resource machine translation task. They built various translation models for roman languages and used these models to evaluate cognate prediction accuracy. Their experiments showed that MT architectures could be successfully used for the cognate prediction task, but not all insights from low-resource MT tasks are applicable and their specificities must be taken into account. (Batsuren et al., 2022) put forward the latest version of an automatically-built database called CogNet. It contains over 8 million cognate pairs in 338 languages. The database was built using an algorithm that employed existing lexicons alongside etymological, orthographic, phonetic, semantic, and geographic features. The CogNet database had a precision score of 95%. (Rani et al., 2023) presented findings from the SIGTYP 2023 shared task on cognate detection for low-resourced languages. The supervised baseline model was a multi-layer LSTM-based network and the unsupervised baseline was a simple Levenshtein edit distance model. The supervised models submitted by the teams performed worse than the baseline. For the unsupervised task, the team used KMeans algorithms to include graphical, phonetic, and language-encoding information and achieved an 11% improvement in accuracy.

4 Corpus

There were two different corpora for the two stages of the experiment. The proof-of-concept was done on a French-Spanish cognate corpus, while for the final realisation a Finnish-Inari Saami corpus was used. Each corpus posed their own challenges, which will be discussed in this section.

4.1 French-Spanish Corpus

The Spanish-French corpus is a cognate data set that has been produced by using cross-lingual embedding methods. It is split into training (3369 possible cognates) and test set (200 gold-standard cognates) and is freely available under the GPL-3.0 license (Sharoff, 2018). For the the work presented in this paper, only the training was used. It was transcribed into the IPA using Epitran[1], a massive multilingual G2P (grapheme to phoneme) system, that offers a systematic method for converting text from various languages into phonetic representations using IPA. While Epitran provides excellent transcriptions for the Spanish words of the corpus, it has only limited support for French due to the language's highly ambiguous orthography. There are plans by the Epitran team to better support French in the future, namely with a different back end model based on Weighted Finite State Transducers or neural networks. The main challenge was thus to detect and correct incorrect transcriptions. A quick scan of the generated transcriptions allowed us to detect recurring errors that we then tried to address with Python code and Regular Expressions. Some of these errors could be explained by the limits of the Epitran module: Incorrect transcriptions of individual letters like <r> in French, or <d> and <g> in Spanish result from an incomplete or improper set of IPA characters in the Epitran mapping files. The French <r> is mapped to the phoneme / ɾ/ instead of /ʁ/, and the Spanish letters <g> and <d> are always mapped to /ɡ/ and /d/ respectively, while they should be transcribed as /ɣ/ and /ð/ in most cases. Other errors, though, could not be explained by wrong mapping, such as the incorrect transcription of the French verbal infinitive ending *-er* as /əʁ/. The preprocessing unit of Epitran should substitute such verb endings by the letter <é> which will then be mapped to /e/, however this seems not to be the case. A possible explanation might be that before the proper rule is applied, another rule is erroneously used where <e>, preceded by a vowel and any character one or more times is transcribed as /ə/, as indicated by this algorithm:

/e/ → /ə/ / (::vowel::).*_

For example, in the verb *accepter* the module would detect that the final <e> is preceded by a vowel and two consonants, which fits the above algorithm and consequently transcribes it as /ə/. These errors have been corrected with Python code and Regular Expressions before the corpus underwent a final, manual check to make sure that all the transcriptions are accurate for training the model.

4.2 Inari Saami-Finnish Corpus

For the final realisation of the project which focused on a marginalized language, a new corpus was compiled. It was decided to use two Uralic languages, specifically Inari Saami and Finnish. The Saami languages are an ensemble of a dozen Uralic languages spoken by Saami people in Northern Europe. A number of them is already extinct or moribund, that is, critically endangered. The Inari Saami community is located in Lapland Province in Finland. To this day the language counts an estimated 300 speakers, which makes it a severely endangered language. To create the corpus, word pairs were taken from the Inari Saami and Finnish word lists from the NorthEuraLex databank (Sammallahti and Morottaja, 1993). Although there has been a noticeable effort in reviving and passing down some of the most widely spoken Saami languages such as Northern Saami (with about 15,000 speakers). Inari Saami has been the subject of much fewer attempts to preserve it. For this reason Inari Saami was chosen as the subject endangered language for this project. Finnish was used (with 5.8 million speakers) as the reference language, due to Finnic languages being traditionally considered the etymologically closest languages to Saami. The parallel Inari Saami - Finnish data of the corpus was extracted from the large-scale lexicostatistical NorthEuraLex database [2], created as a part of the EVOLAEMP Project and available under the CC BY-SA 4.0 international license. It provides lexical data from over 20 language families from Northern Eurasia, with a special focus on Uralic languages. For each language, the database presents a list of English concepts coupled with the related word in the given language, as well as an automatically generated IPA transcription. This latter feature was of particular interest given the nature of this project.

Of the initial 1,558 Inari Saami and 1,188 Finish words, only those were kept that had an equivalent in the respective other language. The remaining

[1] https://github.com/dmort27/epitran/tree/master

[2] http://www.northeuralex.org/ (last access: January 10, 2024)

978 word pairs were then manually checked for phonetic similarity. This was performed in a manner that even the same initial sound was enough to consider the words potential cognates. Nevertheless, soe cognate pairs might have been overlooked, especially in cases where the phonetic relation was not overt. The resulting 500 word pairs were then carefully checked for common Proto-Uralic ancestors. Proto-Uralic is the unattested reconstructed root language of Uralic languages, which includes Finnish and Inari Saami [3]. To this end, the Saami word list on Wiktionary was used, as well as the Álgu database from the Institute for the languages of Finland [4]. The Wiktionary page was assembled based on the data from NorthEuraLex and included several Saami languages along with English, Russian, Finnish, and Hungarian translations [5]. If no information on Proto-Uralic ancestors were available, it was sometimes possible to find other evidence for a cognate relationship between two words, e.g., The Finnish - Inari Saami word pair for *smoke* is *savu* - *suovâ* (orthographic form). While there are no records on a Proto-Uralic root, the Finnish *savu* can be traced back the Proto-Finnic term **savu* which builds a cognate with the Proto-Samic term **suovë*, which again is given as the ancestor of the Inari Saami word *suovâ*. If no evidence could be found to support the claim that two words were cognates, the word pair was removed from the corpus. The remaining 280 word pairs were then completed by 70 word pairs extracted and transcribed manually from a Wiktionary word list of *Proto-Samic terms inherited from Proto-Uralic* [6] as well as guidelines from (Bakró-Nagy et al., 2022) and the NorthEuraLex database.

5 Model

This section provides a quick account of the model and discusses its distinct functions. The model's main function is to predict cognates in a related language using words from a related, reference language while producing a graphic of the models attention in order to aid in explaining the model's predictions.

5.1 Model Architecture

The cognate prediction model is built upon a sequence-to-sequence (seq2seq) neural network, a framework made up of two interacting recursive neural networks (RNNs), known as the encoder and decoder. The model was built using PyTorch, an open-source deep learning module available in Python (Paszke et al., 2017). A seq2seq network's encoder is an RNN that outputs a value for each element in the input phrase. In the case of this model, each element is the integer an IPA character has been mapped to. The encoder then generates a vector and a hidden state for each input element and uses the hidden state for the next part of the input. Furthermore, the decoder is an additional RNN that takes the encoder output vector(s) and generates a series of decoder outputs that, once Softmax is applied to, can be translated into a human-readable-form. Softmax is a function that converts the vector output into a vector of probabilities. In the case of this model, the probability of a character being the next output is found at the index associated with said character.

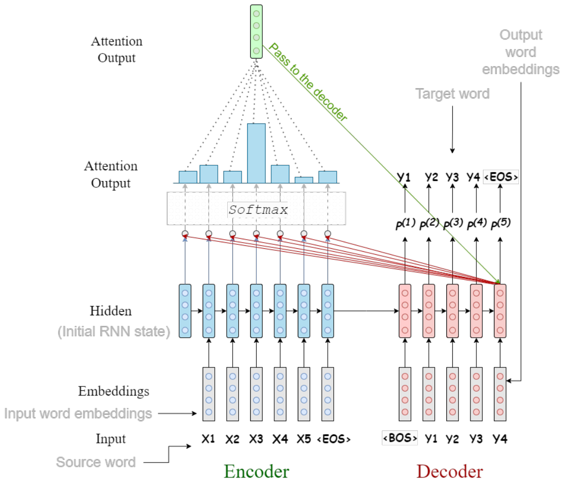

Figure 1: Example of the architecture of the model

This model employs Bahdanau attention (Bahdanau et al., 2016), a feed-forward attention mechanism often employed in such seq2seq models, especially in neural machine translation applications. The input of this layer is the decoder's input and hidden states. A graphic of the architecture can be seen in Figure 1, in which the decoder's inputs are passed to the attention mechanism which then in turn affects the final output of the decoder. The advantage of such attention mechanisms is that

[3] https://en.wikipedia.org/wiki/Proto-Uralic_language (last access: January 9, 2024)

[4] https://kaino.kotus.fi/algu/index.php?t=etusivukkieli=en (last access: January 9, 2024)

[5] https://en.wiktionary.org/wiki/Appendix:Saami_word_lists (last access: January 9, 2024

[6] https://en.wiktionary.org/wiki/Category:Proto-Samic_terms_inherited_from_Proto-Uralic (last access: January 9, 2024

they allow a model to focus on different aspects of the input at each step of the output generation process. This process allows the model to learn the significance of specific elements in the input for making accurate predictions. Importantly, the attention mechanism lets the encoder provide representations for all source tokens, not just the final one. This is achieved by dynamically adjusting context vectors at each step of the input — a departure from older methods using static context vectors. Since the entire process, including the attention function, softmax, and other components, is differentiable, the model may be trained end-to-end without the need for explicit instructions on selecting important input. Attention is calculated during each decoder step by taking into account the decoder state as well as all encoder states, allowing the model to dynamically weigh the importance of different source tokens. The attention output is computed as a weighted sum of encoder states, with attention weights representing the relevance of each source token for the current decoding step. These attention weights are determined through an attention function, producing scores for each encoder state based on its relevance to the current decoder state. A softmax function is then applied to these scores to generate a probability distribution over the source tokens, effectively collecting the model's learned attention at each step.

5.2 Model Training

Both cognate datasets were split 80/20 to create the training and validation datasets respectively. The first model was trained on 2695 pair of French-Spanish that made up the train set and validation loss and performance metrics were calculated after each epoch with the remaining 674 pairs. The second model was trained on the Finnish-Inari dataset. The model was trained on 280 pairs while 70 pairs were held for calculating the validation loss and calculating the metrics. Training of both models was ceased once validation loss did not decrease over the span of five epochs. This is because the current parameters of the model led it to quickly converge and employing this style of early-stopping helps to avoid the problem of overfitting to the training data along with not wasting resources unnecessarily on training that does not lead to substantial improvements to the model. Both models were updated using the loss which was calculated with the negative log likelihood loss (NLL). This metric is used to indicate how close the model's predictions match the expected output. An exhaustive grid search was run to find the optimal parameters for each model. The parameters that were checked were learning rate {0.01, 0.001}, hidden size {128, 512, 1024}, and batch size {8, 32, 64, 128}. The max number of epochs was 1000, in case the early-stopped mechanism never activated.

5.3 Model Evaluation

The models were evaluated using two main metrics. The first of which is the Bilingual Evaluation Understudy (BLEU) score (Papineni et al., 2002). This score measures the similarity of two strings using n-gram similarity along with a brevity penalty (BP) which penalizes the model for making predictions that are shorter than the expected. The n-gram similarity used here is a modified n-gram precision where n-grams of the candidate translation are compared with those in the reference translation adjusting for number of occurrences (p_n). The BLEU score also includes a weight (w_n) for each n-gram size. The metric used to evaluate the model presented in this paper are n-grams of size 1-4 inclusive all equally weighted.

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

The second metric used to evaluate the model is the character n-gram F-score (CHRF). This metric is a character level measure for F-score. F-1 score was used in this paper in order to match that used in articles mentioned later which means that β is 1. The terms $chrP$ and $chrR$ refer to character precision and character recall respectively.

$$\text{chrF}^{\beta} = 100 \cdot \left(1+\beta^2\right) \cdot \frac{\text{chrP} \cdot \text{chrR}}{(\beta^2 \cdot \text{chrP}) + \text{chrR}} \tag{2}$$

6 Results

As stated above a cross-validation grid search was performed in order to find the most optimal parameters for the model. Table 1 displays the best results of this grid search for both the French-Spanish and the Inari-Finnish models. Both models performed most optimally with a learning rate of 0.001 and a hidden size of 512, however the French-Spanish model preferred a batch size of 128 to the Finnish-Inari's 64. The evaluation of the models was performed every epoch as described above and the results are shown in Figure 6 and Figure 7. The best

Dataset	Learning Rate	Hidden Size	Batch Size	Val Loss
French-Spanish	0.001	512	128	0.1189
Finnish-Inari	0.001	512	64	0.4969

Table 1: Best performing parameters obtained from the Grid Search

Paper	F1	BLEU	Architecture
Fourrier		86.6/90.8	SMT/RNN
Batsuren	0.88		Undirected graph
Rani	0.91		Multi-layer LSTM
Model Fre-Spa	.805*	80.2	LSTM with Attention
Model Ina-Fin	0.312*	30.2	LSTM with Attention

Table 2: Comparison of various cognate detection models in BLEU and F1 scores. * Indicates F1 scores that are calculated using CHRF as described in (Popović, 2015)

performing model for the French-Spanish model produced a BLEU score of 80.2 and a CHRF score of .805. The Finnish-Inari model resulted in lower performance with a BLEU score of 30.2 and CHRF score of .312 as the two scores. The training loss for the Finnish-Inari model can be seen to continue decreasing steadily throughout training while the validation loss trended downwards but more sporadically, both seen in Figure 4. The French-Spanish model displays a more uniform decrease in both metrics, as visible in Figure 5.

7 Discussion

The model performed well with French and Spanish but struggled with Inari and Finnish. This drop of performance can be due to different factors. Historical timelines of language divergences can be one of these factors as Finnish and Inari split from their most common ancestor between 1000 BC to 1 AD, Early Proto-Finic (Laakso, 2001a) while French and Spanish have a much more recent common ancestor, diverging in the fifth and sixth centuries AD (Percy, 1887). This could mean that French is closer to Spanish than Finnish is to Inari causing the model to perform worse. Another factor that could have affected the performance of the Finnish-Inari model could be the limited dataset for Finnish and Inari. In fact, the size of the French-Spanish dataset is over 7 times as big. This could have also made it harder for the model to learn the linguistic patterns connecting these two languages, negatively affecting its performance. It is important to note that the performance of the models cannot be directly compared with former attempts which outputted words in the standard orthography of the language as the output of our model is still transcribed in IPA. In a number of languages, including but not limited to English, Finnish, and French, there are more IPA characters that represent the language than characters used in the standard orthography. Our approach stands apart from the methodologies employed in earlier research related to automatic cognate identification, because while former research tested edit distance methods (McCoy and Frank, 2018) or used machine translation approaches (Fourrier et al., 2021) for high-resource languages, the method described in this paper aims to help languages that are under-resourced. Specifically, it focuses on employing machine learning techniques, like LSTM-based networks, on marginalized languages, adapting successful techniques from high-resource languages to address previously overlooked linguistic contexts. In the field of linguistics, the task of cognate prediction/invention relies on theoretical rules linguists discover based on observations of empirical data. Thus, using these rules, linguists can make a laborious effort to prediction of a cognate in a related language, given a word in another. Linguists may examine an etymological database and come up with a set of sound change rules to explain the sound correspondences between the pairs. This also paves the way for the reconstruction of a proto-language and the posited rules are often referred to as reconstruction rules (Meillet, 2020). In the case of Uralic studies, it should be noted that proto-words are transcribed using the Uralic Phonetic Alphabet (UPA). It is the standard notation for the reconstruction rules of Uralic languages. The modern Finnish and Inari Saami words in the dataset are transcribed using IPA. Two different cognate pairs are examined below to demonstrate the process by which linguists find a cognate word's pair in another language. This contrasts with the automated approach to cognate detection presented in this paper, but with the explanation provided through the heat map representation of the model's attention mechanism, aims to aid in understanding the kind of sound change rules the model has been able to identify.

Figure 2 illustrates our first cognate pair to investigate, which is Inari Saami [kod͡zːɐ] to Finnish [kynsi]. Our model has correctly predicted the Finnish pair to be [kynsi]. The Proto-Uralic root

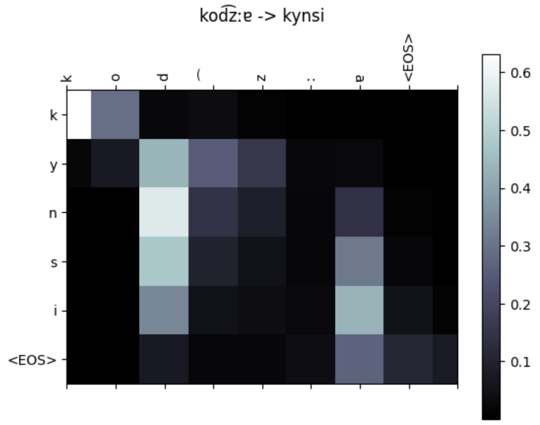

Figure 2: Attention heatmap of [kod͡zːɐ] to [kynsi]

from which these words descend is *künče. The development from Proto-Uralic to Proto-Finnic observes a rule of *č to *t. (Kallio, 2007) *ti is then assibilated to *ci. Assibilation is referred to as a sound change rule in phonology where a consonant turns into its sibilant counterpart. Sibilants are a subset of fricative consonants that are characterized by their higher amplitude and pitch. (Ladefoged and Maddieson, 1996) In further development from Proto-Finnic *künci to modern Finnish, the affricate *c becomes *s. *ü in UPA is equivalent to [y] in IPA. Thus, we end up with Finnish [kynsi]. The historical development from Proto-Uralic to Proto-Saami observes a different set of reconstruction rules. (Sammallahti, 1998) In the Pre-Saami period, *ü turns into *i. Further development into Proto-Saami observes a vowel shift from *i to *ë. Unstressed *e also shows a reflex of *ë, and *č turns into *c as a rule. The development from Proto-Saami *këncë to Inari Saami [kod͡zːɐ] involves other sound change rules. A sound change of the vowel in the first syllable is triggered by the vowel in the second syllable. As a pattern, we observe the development of a first-syllable *ë to *o given another *ë in the second syllable. The unstressed *ë in the second syllable also develops into *a which corresponds to [ɑ] or [ɐ] in Inari Saami based on our observations on the etymological database. Finally, a cluster of a homorganic nasal + a stop or an affricate is merged into the stop/affricate's voiced and geminated counterpart. A homorganic nasal is a nasal stop that has the same place of articulation as the consonant next to it. Gemination refers to the lengthening of a vowel or consonant's duration. Given this rule, merging of the homorganic cluster *nc into the geminated *cc would give us the geminated and voiced alveolar affricate [d͡zː]. For example, given an [o] in Inari Saami, our model has performed a successful transition to a [y] in Finnish. Unlike this automated approach, the comparative reconstruction method takes a different route. The presence of an [o] in Inari Saami could be reconstructed as Inari Saami [o] < Proto-Saami *ë < Pre-Saami *i < Proto-Uralic *ü. This proto-form would, then, be developed as Proto-Uralic *ü > Proto-Finnic *ü > Finnish [y]. The second pair that we will investigate is Inari Saami [polvɐ] to Finnish [pilʋi] as can also be seen in Figure 3. Our model has predicted the Finnish word to be [pulʋi] instead. As per our discussion of the first cognate pair, the unstressed [ɐ] can be reconstructed as Inari Saami [ɐ] < Post-Proto-Saami *a < Proto-Saami *ë < Proto-Uralic *e, and then developed again as Proto-Uralic *e > Proto-Finnic *i > Finnish [i]. The transition performed by the model on the word-final vowel is in line with attested reconstruction rules. This cognate pair descends from the Proto-Uralic word *pilwe. Both Proto-Saami and Proto-Finnic observe a development of Proto-Uralic *w to *v. In the environment of a following vowel, Proto-Finnic *v is realized as the labiodental [ʋ] in Finnish. Throughout the development from Proto-Uralic, *i in the stressed syllable turns into *ë in Proto-Saami. Furthermore, similar to the previous example, *ë in the first-syllable later develops into *o given another *ë in the second syllable. Proto-Finnic does not observe this pattern of sound change; therefore, *i stays the same for Proto-Finnic and Finnish. (Laakso, 2001b) (Lehtinen, 2007) The historical development steps for Proto-Finnic and Proto-Saami can be recounted as follows: Proto-Uralic *pilwe > Proto-Finnic *pilvi > Finnish [pilʋi] and Proto-Uralic *pilwe > Proto-Saami *pëlvë > Post-Proto-Saami *polva > Inari Saami [polvɐ]. In the presence of an [o] in Inari Saami, this time the model has performed a false transition to [u] in Finnish.

8 Carbon Footprint of Model Training

The environmental impact of training machine learning models is a significant aspect of sustainable AI research. In the context of the training of the models presented in this paper, the entire training process, encompassing all pre-final training stages, culminated in a carbon emission of approximately 0.00002 kilograms. This value, was calculated using the Python library CodeCarbon

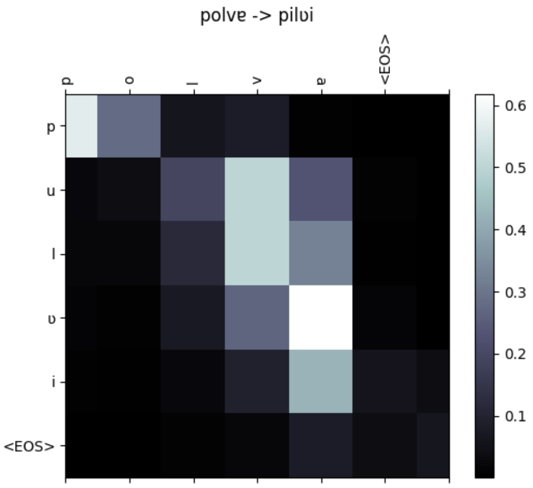

Figure 3: Attention heatmap of [polvɐ] to [pulʋi]

(Lacoste et al., 2019). The carbon footprint of our model while seemingly minuscule, would undoubtedly grow exponentially as more data and training time become necessary. To illustrate, the carbon footprint for the complete training of our model is roughly equivalent to the environmental impact of posting a single tweet. This comparison is founded in an estimate made by Raffi Krikorian, the former Vice President of Engineering at Twitter, now X (Krikorian, 2010). This is only an estimate of the average cost as certain tweets, which attract a larger audience would undoubtedly have an increased cost. There is an even more detailed analysis of carbon emissions associated with various online activities found in (Batmunkh, 2022). In this, one can see that our model, though trained for approximately 8 hours, still incurred a significantly lower carbon cost than consuming social media for the same amount of time. This equivalency underscores the relative efficiency of our training process in terms of its environmental impact. It is prudent to note how surprisingly high, however, the cost of using social media platforms can be and to remain conscious of this while engaging in online activity. This finding aligns with the broader goal of developing environmentally sustainable computational practices. It demonstrates that advanced computational tasks, like training a machine learning model, can be accomplished with minimal environmental costs, akin to commonplace digital actions.

8.1 Implications for Sustainable AI

The minimal carbon footprint of the models' training process not only highlights the efficiency of the training algorithms, but the efficiency of the machine it was trained on. Developing models that are environmentally conscious is crucial for the long-term viability of AI as a field. In fact, the models trained in this paper were not trained on a super computer, rather a low-power Intel mobile processor (Intel Core i5 1155G7). It is also important to note the location in which the training of the model took place, which was in the United States.

- The carbon efficiency of model training must be a key consideration in machine learning and AI development.
- Comparative metrics, like the equivalence to a tweet, make the environmental impact more tangible and understandable.
- Striving for lower carbon footprints in AI aligns with global efforts to reduce greenhouse gas emissions.
- Thinking about where the energy that is used to train a model comes from, to understand its environmental impact.

In conclusion, our approach exemplifies how machine learning can progress in an eco-friendly manner, advocating for a balance between technological advancement and environmental thoughtfulness.

9 Future Work

In order to find more optimized parameters for the training of the model, it may be prudent to use a different search method than grid search. In fact, it has been found that random search can produce comparable or better performing parameters while using a fraction of the computational cost (Bergstra and Bengio, 2012). It would also be interesting to test whether or not the distance of the two languages from their common ancestor, French and Spanish being more closely related than Inari and Finnish, affects performance. This could be done by treating French and Spanish as low resource languages, using only a random subset of comparable size to the corpus of Inari-Finnish for training. Finally, it would be interesting if the model would produce pairs of n-grams that the model believes are strongly related. This could be extracted by getting n-gram frequencies for reference and predicted translations and incorporating the attention vectors for alignment. This sort of output could then be utilized by linguists who are aiming to compare the linguistic changes connecting these languages.

References

Abhaya Agarwal and Jason Adams. 2007. Cognate identification and phylogenetic inference : Search for a better past.

Raimo Anttila. 1989. *Historical and Comparative Linguistics*. John Benjamins.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv:1409.0473 [cs, stat].

Marianne Bakró-Nagy, Johanna Laakso, and Elena Skribnik. 2022. *The Oxford guide to the Uralic languages*. Oxford University Press.

Altanshagai Batmunkh. 2022. Carbon footprint of the most popular social media platforms. *Sustainability*, 14(4).

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2022. A large and evolving cognate database. *Language Resources and Evaluation*, 56(1):165–189.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Clémentine Fourrier, Rachel Bawden, and Benoît Sagot. 2021. Can cognate prediction be modelled as a low-resource machine translation task? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 847–861, Online. Association for Computational Linguistics.

David Hall and Dan Klein. 2010. Finding cognate groups using phylogenies. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1030–1039.

Petri Kallio. 2007. *Kantasuomen konsonanttihistoriaa*, Suomalais-ugrilaisen seuran toimituksia, pages 229–249. Suomalais-Ugrilainen Seura, Suomi.

Raffi Krikorian. 2010. From chirp: Energy / tweet 100 j something / tweet http://post.ly/ae6f.

Johanna Laakso. 2001a. The Finnic languages. In Östen Dahl and Maria Koptjevskaja-Tamm, editors, *Circum-Baltic Languages: Volume 1: Past and Present*, Studies in Language Companion Series, pages clxxix–ccxii. John Benjamins Publishing Company.

Johanna Laakso. 2001b. *The Finnic languages*, pages clxxix–ccxii.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *CoRR*, abs/1910.09700.

P. Ladefoged and I. Maddieson. 1996. *The Sounds of the World's Languages*. Phonological Theory. Wiley.

Tapani Lehtinen. 2007. *Kielen vuosituhannet: suomen kielen kehitys kantauralista varhaissuomeen*. Number 215 in Tietolipas. Suomalaisen Kirjallisuuden Seura, Suomi.

Richard T McCoy and Robert Frank. 2018. Phonologically informed edit distance algorithms for word alignment with low-resource languages. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2018*, pages 102–112.

A. Meillet. 2020. *La méthode comparative en linguistique historique*. Prodinnova.

Carlo Meloni, Shauli Ravfogel, and Yoav Goldberg. 2021. Ab antiquo: Neural proto-language reconstruction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4460–4473, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Thomas Percy. 1887. *Reliques of Ancient English Poetry: Consisting of Old Heroic Ballads, Songs, Etc*. F. Warne and Company. Google-Books-ID: RPg5AAAAMAAJ.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Priya Rani, Koustava Goswami, Adrian Doyle, Theodorus Fransen, Bernardo Stearns, and John P. McCrae. 2023. Findings of the SIGTYP 2023 shared task on cognate and derivative detection for low-resourced languages. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 126–131, Dubrovnik, Croatia. Association for Computational Linguistics.

Pekka Sammallahti. 1998. *The Saami languages: an introduction*. Davvi girji, Kárášjohka.

Pekka Sammallahti and Matti Morottaja. 1993. *Säämi-suomâ sänikirje - Inarinsaamelais-suomalainen sanakirja*. Girjegiisá Oy, Ohcejohka.

Serge Sharoff. 2018. Language adaptation experiments via cross-lingual embeddings for related languages.

In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Appendices

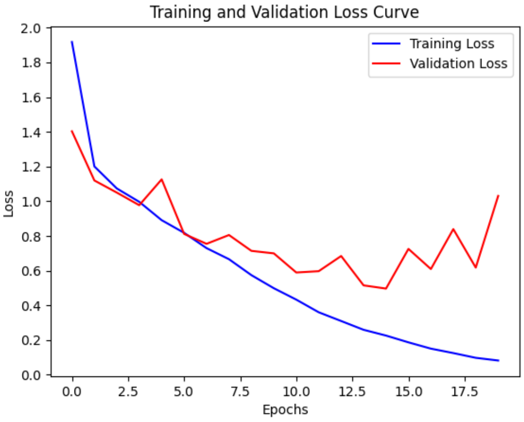

Figure 4: Training and Validation Loss over Epochs when training on the Finnish-Inari Corpus

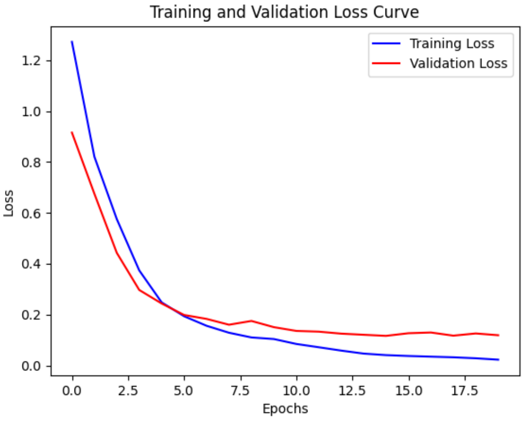

Figure 5: Training and Validation Loss over Epochs when training on the French-Spanish Corpus

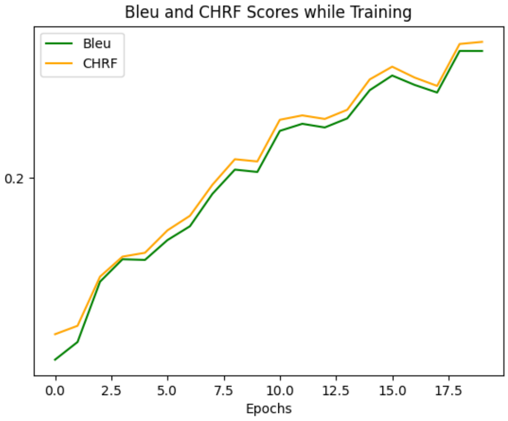

Figure 6: CHRF and BLEU score performance while training on the Finnish-Inari Corpus

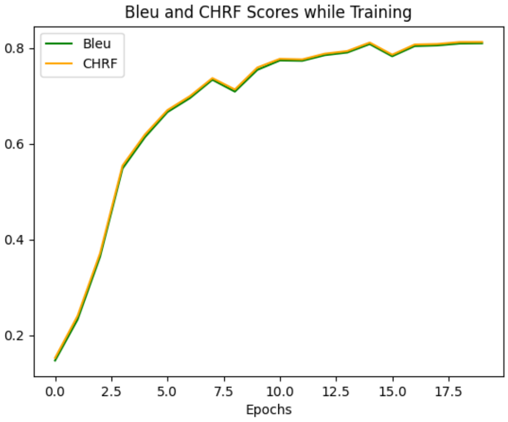

Figure 7: CHRF and BLEU score performance while training on the French-Spanish Corpus