– Independent Work Report Spring 2020 –

# Deep Learning for Cognate Identification

Eve Fleisig
Adviser: Christiane Fellbaum

## Abstract

*Identifying cognates–words with a shared origin across languages–aids researchers in fields from historical linguistics, where cognates help to reconstruct the relationships between languages, to machine translation, where cognates augment systems for translating low-resource languages to and from high-resource relatives. However, most approaches for identifying cognates automatically have sought to improve on metrics that measure character overlap, such as Levenshtein distance or pointwise mutual information. Minimal work has been done on using neural networks to identify cognates, nor has any approach taken advantage of deep character-level models to extract linguistic information beyond character correspondences. This paper presents a neural model for cognate identification that incorporates information from a character-level CNN language model trained on cognate pairs. It finds that the model correctly identifies cognates in a dataset of Romance and Germanic word pairs that have undergone a variety of phonological, orthographic, and morphological changes. The model also successfully transfers this knowledge to the task of identifying cognates between Spanish and Basque, a previously unseen low-resource language. The results suggest that the model learns characteristics of cognates in general from the larger dataset and can apply them to a new language pair. They also suggest that efforts to identify cognates in low-resource languages are significantly aided by pre-training on larger cognate datasets from higher-resource languages before fine-tuning on the limited data available for the low-resource language.*

# 1. Introduction

The value of identifying cognates–words with a shared origin across languages–stretches from reconstruction of ancient languages to improving state-of-the-art machine translation systems. In historical linguistics, identifying cognates between languages is necessary for reconstructing the relationships between languages, but manually-constructed cognate lists do not exist for many language pairs. Natural language processing researchers seeking to translate low-resource languages to and from high-resource relatives, particularly useful for translating material between a country's majority language and minority languages, face the difficulty of training with scarce data. By augmenting models with information about cognates between the two languages, researchers have been able to significantly improve machine translation systems' performance on such language pairs.

These fields benefit from the ability to identify cognates automatically. However, most approaches for identifying cognates have sought to improve on metrics that measure character overlap, such as Levenshtein distance or pointwise mutual information. Minimal work has been done on using neural networks to identify cognates, nor has any approach taken advantage of deep character-level models to extract linguistic information beyond character correspondences.

This paper presents a neural model for cognate identification that incorporates information from a character-level CNN language model trained on cognate pairs. The model identifies cognates in a dataset of Romance and Germanic word pairs and successfully transfers this knowledge to the task of identifying cognates between Spanish and Basque, a previously unseen low-resource language.

Section 2 gives an overview of the context of this problem and previous work on cognate identification, while Section 3 describes the motivation for the approach in this work. Section 4 details the process of training and evaluating the cognate classification model. Section 5 analyzes the overall performance of the models, the performance on individual language families, and the performance of transfer learning on the Basque dataset. Section 6 examines the model's ability to identify cognates in both datasets with different types of phonological, orthographic, or morphological correspondences.

# 2. Background and Related Work

## 2.1. Background

Computational approaches using cognates to measure the similarity between different languages date to the mid-1900s. Morris Swadesh proposed the use of glottochronology, in which the number of cognates between two languages for a relatively stable set of words over time is compared to determine the relationship between the languages and the time of their divergence [18]. Though later linguists challenged Swadesh's assumptions, the principle of identifying cognates to reconstruct phylogenetic trees of the relationships between languages has remained fundamental to the comparative method in historical linguistics, as modern linguists use manual comparison or metrics such as substring similarity to quantify the similarity between known cognates [1]. Identifying systematic sound correspondences between known cognates thus allows linguists to trace the descent of languages.

The advent of natural language processing highlighted the importance of cognate identification not only in historical linguistics, but also in language pedagogy and machine translation. Because historical linguists rely on identifying correspondences between cognates in languages for which cognate lists may not exist, automatic tools for identifying potential cognates may save these researchers significant work [9]. In addition, second language learners greatly benefit from awareness of cognates between their native language and the language they are acquiring [21]. Since lists of manually identified cognates are not readily available for all language pairs, automatic cognate identification can assist in providing cognate lists to language learners.

In machine translation, a prevalent challenge is the translation of low-resource languages with high-resource relatives. In countries such as the United Kingdom, Spain, and Finland, translating material from the majority language to minority languages is crucial for communicating with those who may not speak the majority language and for revitalizing endangered minority languages. However, minority languages often lack the extensive written material necessary for training machine translation systems, even when the language is closely related to or borrows extensively from a high-resource relative.

Recent studies have found that augmenting machine translation systems with information about cognates between a low-resource language and a high-resource relative significantly improves models' performance [6, 13]. However, these methods require documented lists of cognates between the low-resource and high-resource languages, information that is similarly rare. Automatic cognate identification provides an avenue for detecting the cognates needed to improve low- to high-resource machine translation systems when cognate lists are unavailable.

### 2.2. Related Work

Until recently, most approaches to automatic cognate detection drew on character correspondences and basic metrics for sequence alignment and edit distance. Most metrics were based on Levenshtein distance, which measures the minimum number of single-character edits (insertion, deletion, or substitution) needed to turn a source string into a target string [8]. To improve on this metric, Rama et al. (2015) used subsequence similarity metrics [14], while Jäger et al. (2017) and Rama et al. (2017) used pointwise mutual information[1] as a measure of similarity [5, 15]. These approaches all used support vector machines to classify words as cognates or non-cognates. Other research by Beinborn et al. [2] and Malmasi et al. [10] approached the problem as a character-level statistical machine translation problem.

Most recently, Soisalon-Soininen and Granroth-Wilding (2019) presented draft work using a Siamese convolutional neural network (S-CNN) to identify cognates [17]. In their preliminary results, the S-CNN outperforms models using Levenshtein distance and SVM models. For each word in a pair $(a,b)$, input feature vectors whose columns represent one-hot encoded characters are convolved with a common filter $W$ to produce vectors $x_a$ and $x_b$. After undergoing max-pooling, the vector difference between $x_a$ and $x_b$ passes through a fully-connected layer.

[1] $PMI(a,b). = \frac{logs(a,b)}{q(a)q(b)}$, where $a,b$ are characters in phonetic transcription, $s(a,b)$ is the probability of $a$ and $b$ being aligned to each other in a pair of cognate words, and $q(a),q(b)$ are the probabilities of occurrence of $a$ and $b$. Positive PMI scores provide evidence for cognacy [5].

Notably, Soisalon-Soininen and Granroth-Wilding's approach lacks several useful features for determining the cognacy of words. First, the model is unaware of the languages that the words are from, information available in the initial word lists that makes it possible to train the model to identify language-specific sound correspondences. Secondly, because the input features are one-hot vectors encoding individual characters, the model does not take any information into account beyond character correspondences, particularly information about the structure of the individual languages that can be gathered from a language model. Finally, by training the hidden layer only on the *difference* between the filtered vectors $x_a$ and $x_b$, the model loses information about potential correspondences that could be conserved by giving the hidden layer access to information about $x_a$ and $x_b$ themselves.

Outside of cognate identification, however, research on character-level neural models suggests that these models may be well-suited to modeling cognate pairs. Drawing on Word2vec [11], which uses a neural network to create vector embeddings for words based on their contexts, new research has resulted in methods for creating vector embeddings suited to working with individual words. These include morph2vec, which models words' morphemes [20], and chars2vec, which models individual characters [3]. Kim et al. (2016) created a neural language model whose inputs, rather than word embeddings, are based on subword units: the outputs of a single-layer character-level convolutional neural network (detailed in Section 4.4). Though not yet applied to cognate identification, this approach has improved models for similar tasks, such as neural machine translation of rare words [16].

## 3. Approach

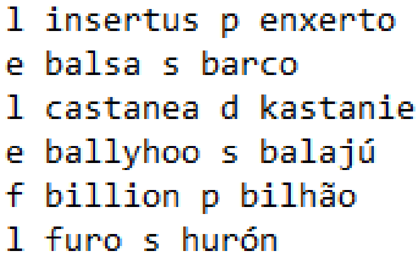
Given the gap in the literature on neural approaches to cognate identification and the limitations of the minimal existing research, this project presents several innovations for cognate identification. First, it presents concrete results on the success of neural networks at identifying cognates. Second, it incorporates information from a character-level CNN language model trained on cognate pairs to strengthen the model beyond simply identifying character correspondences. Finally, it finds that this model is able to identify cognate pairs between a known high-resource language and an unseen low-resource one, particularly useful for machine translation.

# 4. Implementation

## 4.1. Overview

The implementation proceeded by first training a character-level CNN language model on a dataset of Romance and Germanic cognate pairs. Then, for each pair of words in the training set (half cognates, half non-cognates), the language model's encoding of the word pair is concatenated to a vector of one-hot encoded characters in the word pair. A 2-layer neural network is trained to classify these vectors as cognates or non-cognates, then tested on new Romance and Germanic word pairs and Basque-Spanish word pairs. When tested on the Basque-Spanish word pairs, three versions of the model were tested: (1) trained on the Romance and Germanic data alone, (2) pre-trained on the Romance and Germanic data and fine-tuned on the Basque data, and (3) trained on the Basque data alone.

## 4.2. Datasets



```
l insertus p enxerto
e balsa s barco
l castanea d kastanie
e ballyhoo s balajú
f billion p bilhão
l furo s hurón
```

**Figure 1: Sample cognate entries from the dataset of Romance and Germanic cognates.**

The datasets for the model, a Romance and Germanic language dataset and a Basque-Spanish dataset, were chosen for their relevance to the different applications of cognate identification.

The Romance and Germanic language dataset is aimed at the task of identifying cognates between related languages. Results on the Romance languages alone and the Germanic languages alone were also examined for the application of identifying cognates between closely related languages.

|  | Spanish | French | Italian | Port. | Latin | German | Dutch | English | Danish | Swedish |
|---|---|---|---|---|---|---|---|---|---|---|
| Spanish | - | 656 | 188 | 122 | 7022 | 50 | 16 | 1012 | 22 | 2 |
| French | 656 | - | 788 | 324 | 5750 | 1646 | 598 | 5518 | 568 | 140 |
| Italian | 188 | 788 | - | 110 | 6932 | 152 | 26 | 826 | 60 | 8 |
| Port. | 122 | 324 | 110 | - | 3662 | 34 | 8 | 354 | 6 | 0 |
| Latin | 7022 | 5750 | 6932 | 3662 | - | 572 | 316 | 13066 | 494 | 160 |
| German | 50 | 1646 | 152 | 34 | 572 | - | 70 | 888 | 294 | 78 |
| Dutch | 16 | 598 | 26 | 8 | 316 | 70 | - | 434 | 30 | 8 |
| English | 1012 | 5518 | 826 | 354 | 13066 | 888 | 434 | - | 236 | 160 |
| Danish | 22 | 568 | 60 | 6 | 494 | 294 | 30 | 236 | - | 46 |
| Swedish | 2 | 140 | 8 | 0 | 160 | 78 | 8 | 160 | 46 | - |

**Table 1: Number of cognates for each language pair in the Romance-Germanic dataset.**

The Spanish-Basque dataset targets the issue of identifying cognates between languages with a high demand for translation for which one language has significantly fewer resources than the other. Basque, a language isolate spoken in northern Spain, is a minority language that has recently undergone extensive revitalization. The Basque government has accompanied these efforts with investment in Basque machine translation. However, research has been hindered by the scarcity of Basque translation data, such that even neural models like Google Translate struggle to grasp Basque morphology [19]. Despite being genetically unrelated, Basque vocabulary has significant numbers of cognates with Spanish words due to borrowings from Spanish, Latin, French, and other Romance languages, making the identification of cognates useful for augmenting machine translation systems.

The Romance-Germanic dataset was taken from Etymological WordNet [4] and contains pairs of cognates between the five Romance and five Germanic languages with the most existing data: Spanish, French, Italian, Portuguese, Latin, German, Dutch, English, Danish, and Swedish. The Romance and Germanic families were chosen because the most data was available for these families and to minimize the issues posed by different character sets (e.g., Cyrillic) used in other language families. A total of 53,452 cognate pairs were used; Table 1 displays the number of cognates for each language pair. Due to the lack of extensive Basque-Spanish cognate lists, the dataset of 1000 Basque-Spanish cognates[2] was gathered manually from the Etymological Dictionary of Basque

[2] The resulting Basque-Spanish cognate list is available for public use on GitHub.

[12].

For both datasets, cognates may be inherited directly by one language from the other or may stem from a shared parent language. The cognates may have different definitions so long as they share the same etymological origin. To train the model to distinguish cognate pairs from non-cognate pairs and control for the different numbers of cognates among different language pairs, for every cognate pair $a, b$ between languages $L_1$ and $L_2$, a non-cognate pair $c, d$ was added to the dataset, where $c \in L_1$ and $d \in L_2$ are randomly sampled words that are not cognates. Thus, given a word pair $x, y$ in the dataset from a language pair $L_1, L_2$, a random guess has a 50% chance of a correct classification.

All word pairs are of the form $l_1$ $w_1$ $l_2$ $w_2$ separated by delimiters, where $l_1$ is a single character representing the language of word $w_1$ and $l_2$ represents the language of word $w_2$. Each dataset was divided into training, development, and test sets using a 70/20/10 split.

### 4.3. Model

Figure 2 depicts the structure of the cognate classification model. After creating the datasets, a character-level CNN language model is trained on Romance and Germanic cognate pairs. Then, for each pair of words in the training set, the language model's encoding of the word pair is concatenated to a vector of one-hot encoded characters in the word pair. A 2-layer neural network is trained to classify these vectors as cognates or non-cognates.

All code and models can be found on GitHub at `github.com/efleisig/cognate_nn_iw`.

#### 4.3.1. The Character-Level CNN Language Model

To train a language model to recognize cognate pairs, the CharCNN model from Kim et al. (2016) is trained on the Romance and Germanic cognates in the training set (Figure 3). For each input word $w$, CharCNN combines the word with the input history, represented by the hidden state, to predict the next word. The first layer creates a matrix $C_k$ of stacked character embeddings. Then, $C_k$ is convoluted with multiple filter matrices examining substrings of different lengths that aim to capture the most important features within different subword units.

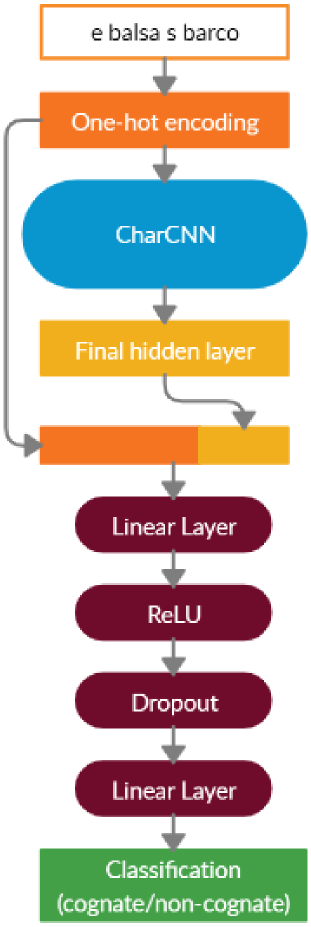


**Figure 2: Structure of the cognate classification model.**

Applying max-over-time pooling yields a fixed-dimensional representation $y$ of the word, which is given to the highway network. The highway network, similar to long-short term memory cells, mitigates the effects of vanishing gradients on deep networks with layers structured as follows:

$$z = t \odot g(W_H y + b_H) + (1 - t) \odot y$$

where $y$ is the input vector, $g$ is a nonlinearity, the transform gate $t = \sigma(W_T y + b_T)$, and the carry gate is $(1 - t)$. This allows for training the deep network by adaptively carrying some input dimensions directly to the output.

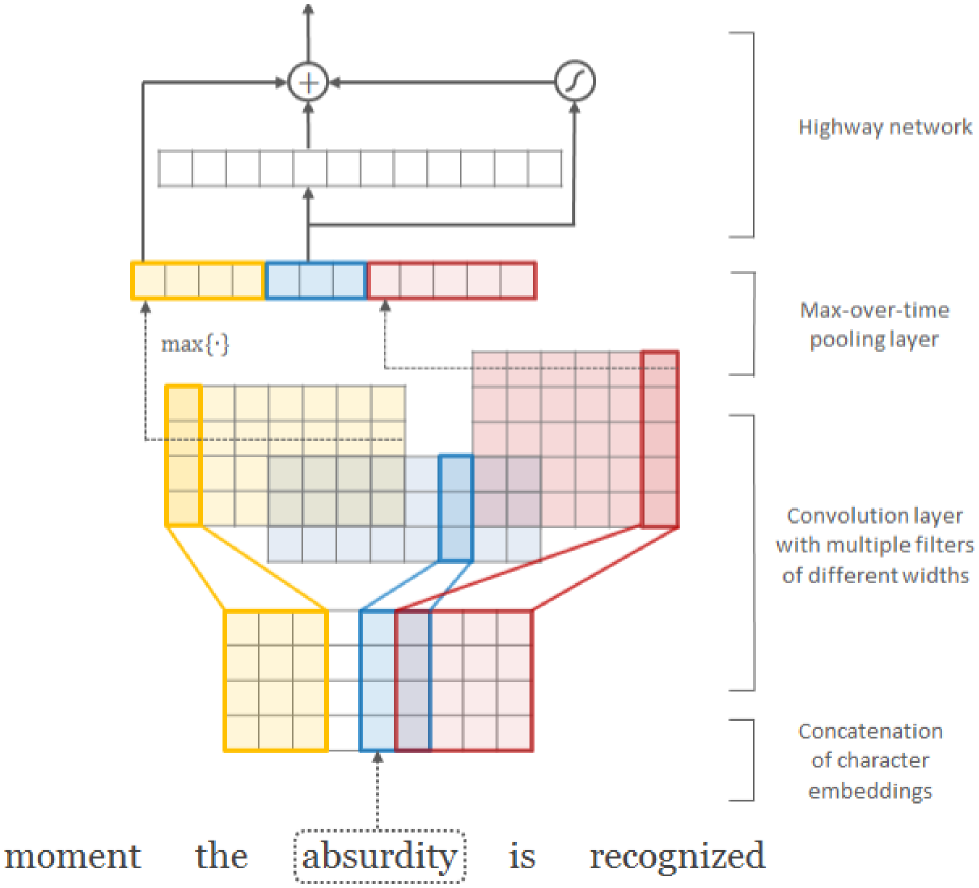

Highway network
Max-over-time pooling layer
max{·}
Convolution layer with multiple filters of different widths
Concatenation of character embeddings
moment
the
absurdity
is
recognized

**Figure 3: Structure of the CharCNN model from Kim et al. (2016). For this model, the input phrase "moment the absurdity is recognized" is replaced by a cognate entry, such as "e balsa s barco".**

### 4.4. Classification

The representation of a word pair from the last hidden layer of CharCNN undergoes max-pooling to reduce its dimensionality and is concatenated with the one-hot representation of the word pair to form the vector **x**. **x** is passed to the classifier, which consists of a hidden linear layer of size 100 with ReLU activation, a dropout layer with 40% chance of dropout,[3] and a second linear layer to output **y**, which scores the probability that a pair belongs to the class of cognates or non-cognates:

[3] Dropout randomly zeroes some elements to avoid overfitting.

$$\mathbf{z_1} = \mathbf{x}\mathbf{W_1}^T + \mathbf{b_1} \qquad \text{(First linear layer)} \tag{1}$$

$$\mathbf{z_2} = max(0, \mathbf{z_1}) \qquad \text{(ReLU)} \tag{2}$$

$$\mathbf{z_3} = Dropout(\mathbf{z_2}, 0.4) \tag{3}$$

$$\mathbf{y} = \mathbf{z_3}\mathbf{W_2}^T + \mathbf{b_2} \qquad \text{(Second linear layer)} \tag{4}$$

where $W_1, W_2$ and $b_1, b_2$ are trained weights and biases, respectively.

The model is trained on the Romance-Germanic dataset and tested on the Romance-Germanic dataset and the Basque-Spanish dataset, as detailed below.

## 5. Results and Evaluation

Section 5.1 presents the overall results on the Romance and Germanic dataset, Section 5.2 examines analyzes the results by language family, and Section 5.3 presents the results on the Basque-Spanish dataset. Section 6 presents examples of the models' ability to capture different types of correspondences between cognates.

### 5.1. Performance on the Romance-Germanic dataset

After testing the classification model on the Romance-Germanic dataset, the accuracy, precision, recall, and F1 scores[4] of the model were calculated. The model was compared to a baseline support vector machine (SVM) model, as that was the model type used by Jäger et al. (2017) and Rama et al.

[4] Precision:

$$\frac{\text{[pairs marked cognates that are cognates]}}{\text{[pairs marked cognates that are cognates]+[pairs marked cognates that are not cognates]}}$$

Recall:

$$\frac{\text{[pairs marked cognates that are cognates]}}{\text{[pairs marked cognates that are cognates]+[pairs marked not cognates that are cognates]}}$$

F1 Score:

$$\frac{\text{2*Precision*Recall}}{\text{Precision + Recall}}$$

(2017) in relatively successful non-neural studies, trained on the number of indexes with matching characters in both words.[5]

For all four metrics, the neural model outperformed the baseline (Table 2), with an accuracy of 95.6%, F1 score of 0.957, precision of 0.944, and recall of 0.971.[6]

| Model | Accuracy (%) | F1 | Precision | Recall |
|---|---|---|---|---|
| NN | **95.6** | **0.957** | **0.944** | **0.971** |
| Baseline (SVM) | 86.1 | 0.869 | 0.822 | 0.922 |

**Table 2: Performance of the neural model, compared with the SVM baseline, on Romance and Germanic cognate identification.**

## 5.2. Performance by Language Family

The accuracy, precision, recall, and F1 scores were also calculated for the subgroups of the data by language family: Romance words paired with Romance words, Germanic words paired with Germanic words, and Romance words paired with Germanic words.

The model performs moderately better at classifying Romance and Germanic words paired together. One explanation for this difference may be that the model performs better at identifying true cognates when the structure of the languages and the characters they use is already fairly different, such that true cognates stand out more from non-cognates.[7]

| Family | Accuracy | F1 | Precision | Recall |
|---|---|---|---|---|
| Romance-Romance pairs | 94.9 | 0.950 | 0.940 | 0.960 |
| Germanic-Germanic pairs | 94.2 | 0.943 | 0.934 | 0.950 |
| Romance-Germanic pairs | 96.1 | 0.964 | 0.948 | 0.982 |

**Table 3: Performance of the neural model on Romance and Germanic cognate identification.**

[5] This method, which requires manually selecting features, outperformed training the SVM on the one-hot vectors used for the neural model, for which the results were no better than random.

[6] Also for comparison, a random baseline (or guessing the majority class, since the number of cognates and non-cognates is even), would have an accuracy of 50% and F1 score of 0.5.

[7] The smaller difference in performance between the Romance-Romance and Germanic-Germanic cognates may be due to the fact that the dataset contained more Romance-Romance cognates than Germanic-Germanic cognates (although the number of Romance-Romance and Romance-Germanic cognates is roughly equal).

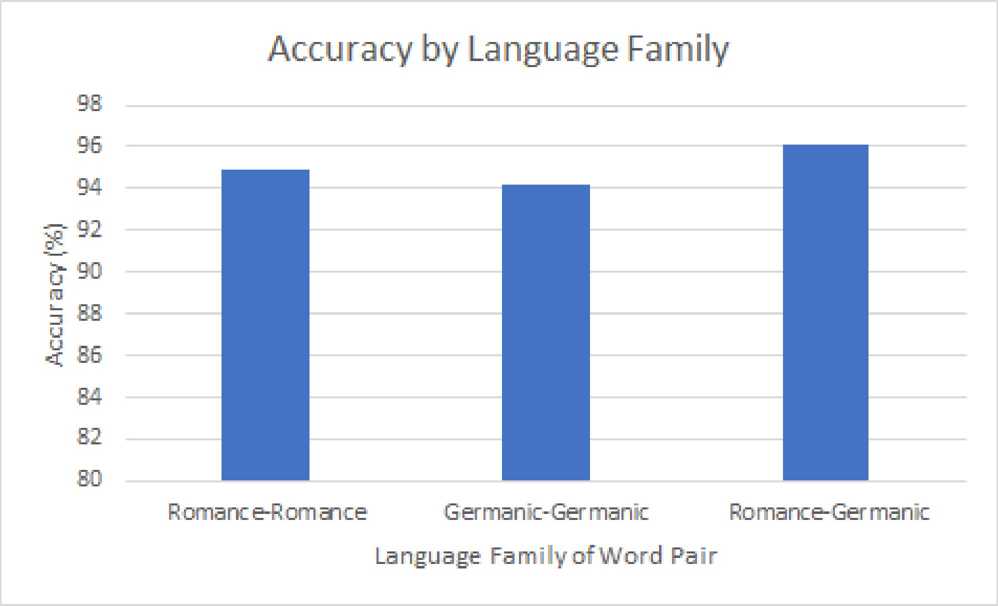


**Figure 4: Classification accuracy for pairs of Romance words paired with each other, Germanic words paired with each other, and Romance words paired with Germanic words.**

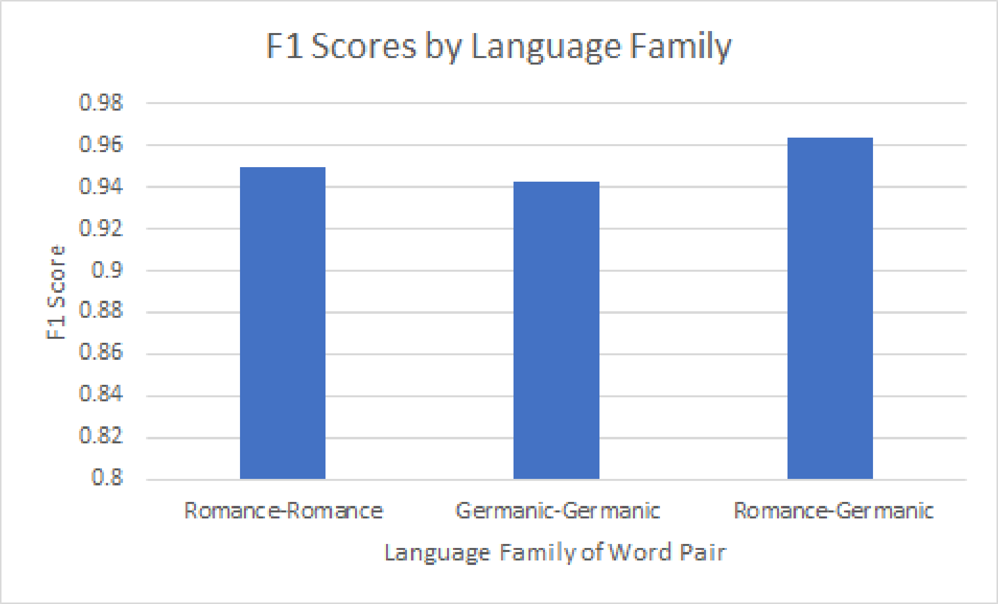


**Figure 5: Classification F1 scores for pairs of Romance words paired with each other, Germanic words paired with each other, and Romance words paired with Germanic words.**

### 5.3. Performance on the Basque-Spanish Dataset

To determine the model's capacity for transfer learning, it was then tested on the dataset of Basque and Spanish cognates. The model's ability to generalize onto a new language that it has not seen before–and, moreover, from a language family not seen before–would indicate its usefulness for identifying cognates between high- and low-resource languages and therefore its applicability to machine translation.

The model was tested on the Basque dataset under three conditions: directly classifying Basque-Spanish cognates after pre-training on the Romance-Germanic dataset, fine-tuning on the smaller Basque-Spanish training set before classification (under 2% of the Romance-Germanic dataset's size), and training the model only on the Basque-Spanish data. This performance was compared to that of the SVM when trained on the Romance-Germanic and Basque-Spanish data, on the Romance-Germanic data alone, and on the Basque-Spanish data alone.

The neural model outperformed the baseline, with higher accuracy and F1 scores under all three conditions than the best SVM model (Table 4). The neural model performed best when pre-trained on the Romance-Germanic dataset and fine-tuned on the Basque dataset (86.3% accuracy, 0.869 F1). However, even the pretrained model without Basque fine-tuning (84.0% accuracy, 0.828 F1) performed better than the model trained on the Basque data alone (76.5% accuracy, 0.791 F1).

The improvement in performance from pre-training on the Romance-Germanic dataset suggests that the model is able to learn characteristics of cognates in general from the Romance-Germanic dataset and apply them to a new language pair. In addition, it suggests that efforts to identify cognates in low-resource languages are significantly aided by pre-training on cognate datasets from higher-resource languages for which more data is available.

| Model | Accuracy | F1 |
|---|---|---|
| NN, fine-tuned (a) | **86.3** | **0.869** |
| NN, no fine-tuning (b) | 84.0 | 0.828 |
| NN, no pretraining (c) | 76.5 | 0.791 |
| SVM baseline (RG + Basque training) (a) | 71.2 | 0.761 |
| SVM baseline (RG training only) (b) | 71.2 | 0.761 |
| SVM baseline (Basque training only) (c) | 74.4 | 0.658 |

**Table 4: Performance of the neural model (NN), compared with the SVM baseline, on Basque cognate identification. Each model was tested by (a) training on the Romance-Germanic dataset and testing directly on Basque cognates, (b) training on the Romance-Germanic dataset and fine-tuning on the Basque dataset, and (c) training only on the Basque dataset.**

## 6. Qualitative Analysis of Specific Types of Change

### 6.1. The Romance-Germanic Dataset

Examining the model's classification of the test set pairs indicates that the model is able to identify pairs of cognates with different types of character correspondences resulting from different types of phonological, orthographic, or morphological changes.

In cognates from a mother language and a daughter language, the dataset contains examples of multiple types of sound change (Table 5). For example, elision, the omission of sounds, occurs in pairs such as Latin *insula* and Portuguese *ilha*. Prothesis, the addition of a sound at the beginning of a word, occurs in cognate pairs such as Latin *scutella* and Spanish *escudilla* (following the regular epenthesis of */e/* before word-initial clusters that begin with */s/*). Metathesis, the transposition of sounds, occurs in pairs such as Italian *cagliari* and Latin *caralis*. A future extension to this research could be to determine whether the model is able to identify the specific types of sound changes occurring between cognates in a mother language and a daughter language.

In words with less direct relationships than mother and daughter languages, the model must also capture regular character correspondences stemming from phonological or orthographic changes. The German word *Dschunke* and Portuguese *junco* have only two characters in common, requiring the model to capture the correspondences between the consonant clusters *dsch* and *j* and between *k* and *c*. Likewise, the Latin *vicis* and Portuguese *vez* have only one character in common, requiring the model to capture the correspondences between the vowels *i* and *e* and between the consonants *c* and *z*. Cognates with few characters in common pose a challenge for metrics like Levenshtein distance that focus on the number of differing characters. However, because cognates with few characters in common generally obey regular sound correspondences, the ability to capture these correspondences likely allows the model to surpass metrics like the Levenshtein distance.

| Word Pair | Type of Change |
|---|---|
| ilha (Portuguese) and insula (Latin) | Elision |
| escudilla (Spanish) and scutella (Latin) | Prothesis |
| cagliari (Italian) and caralis (Latin) | Metathesis |
| Dschunke (German) and junco (Portuguese) | Consonant correspondences |
| vez (Portuguese) and vicis (Latin) | Vowel correspondences |

**Table 5: Examples of different types of sound changes and character correspondences in correctly classified cognates.**

Examining the pairs that the model classified incorrectly suggests that most errors occurred on cognate pairs with lower character overlap and for which fewer pairs in the dataset had similar sound correspondences. For example, the model did not recognize the French *geindre* and Latin *iunior* as cognates; no other French words contain the trigram *gei* and no other Latin words beginning with *i* have French cognates beginning with *g*.

### 6.2. Specific Types of Change in the Basque-Spanish Dataset

Basque-Spanish cognates present character correspondences not seen in the Romance-Germanic dataset stemming from several phonological and orthographic changes into Basque that do not occur in the Romance-Germanic examples (Table 6). For example, the Spanish word *terco* underwent sonorization (*/t/ > /d/), an orthographic change (*⟨*c*⟩ > ⟨*k*⟩), and metathesis (*derko* > *dekor*) before becoming the Basque word *dekor*. The Spanish *emplear* and Basque *enplegatu* share the common Gascon root *emplegà*, but whereas the Spanish word underwent elision (*g > ∅ / V_V ) and changed the verb ending to the regular Spanish suffix *-ar*, the Basque word underwent dissimilation (*m > n / _b,p) and changed the ending to the regular Basque suffix *-atu*.

| Spanish Word | Basque Word | Earlier Form | Changes |
|---|---|---|---|
| *chico* | *txiko* | Spanish *chico* | Orthographic change (*⟨*c*⟩ > ⟨*k*⟩)<br>Orthographic change (*⟨*ch*⟩ > ⟨*tx*⟩) |
| *almidón* | *amirun* | Spanish *almidón* | Lenition (*/d/ > /ɾ/)<br>Raising (*/o/ > /u/) |
| *pavés* | *babes* | Spanish *pavés* | Merger (*/p/, /v/ > /b/) |
| *terco* | *dekor* | Spanish *terco* | Sonorization (*/t/ > /d/)<br>Orthographic change (*⟨*c*⟩ > ⟨*k*⟩)<br>Metathesis (*derko* > *dekor*) |
| *acabar* | *bukatu* | Spanish *acabar* | Metathesis (*acabar* > *abacar*)<br>Orthographic change (*⟨*c*⟩ > ⟨*k*⟩)<br>Morphological change (*-ar* > *-atu*) |
| *emplear* | *enplegatu* | Gascon *emplegà* | Basque: Dissimilation (*m > n / _b,p)<br>Basque: Morphological change (*-à* > *-atu*)<br>Spanish: Elision (*g > ∅ / V_V )<br>Spanish: Morphological change (*-à* > *-ar*) |

**Table 6: Examples of phonological, orthographic, and morphological differences between Spanish and Basque in the dataset.**

In some word pairs, particularly those derived not from Spanish directly but from a shared language of origin, these changes can result in words with very few characters in common (either in corresponding positions or in the words as a whole), as seen in Table 7. For example, the Spanish word *pez* and Basque *bike* have no matching characters in corresponding positions, but both share the common Latin root *pice(m)*.

| Root | Spanish Word | Basque Word | Matching Characters in Corresponding Positions | Shared Characters |
|---|---|---|---|---|
| Latin *pice(m)* | *pez* | *bike* | 0 | 1 |
| Latin *cochlear* | *cuchara* | *goilare* | 2 | 2 |
| Latin *filum* | *hilo* | *firu* | 1 | 1 |
| Proto-Romance *cimis* | *chinche* | *zimitz* | 1 | 1 |

**Table 7: Examples of Spanish and Basque words from a shared root with minimal character overlap.**

Moreover, some morphological changes result in character correspondences unique to the Basque dataset. For example, the Basque cognates contain significantly more examples of reduplication than the Romance-Germanic ones. Used to communicate intensity, iterativity, or focalization, Basque reduplication is a productive morphological process (such as in *aitaita* "grandfather", from *aita* "father") and results in words with lower resemblance to Spanish cognates [7]. In the dataset, this results in cognates such as the Spanish *dudar* and Basque *durduratu*.

As with the Romance-Germanic dataset, errors tended to occur on words with low character overlap and whose character correspondences occurred less frequently in the dataset. However, fine-tuning appeared particularly useful in allowing the model to identify correspondences specific to Basque-Spanish word pairs. Without fine-tuning, the model did not identify Basque *durduratu* and Spanish *dudar* as cognates, nor Basque *zimitz* and Spanish *chinche*, but the fine-tuned model was able to identify both as cognates. This difference suggests that exposure to a comparatively small amount of Basque data allowed the model to recognize word pairs as cognates when the word pairs contained language-specific morphological changes (e.g., reduplication) and language-specific phonological and orthographic correspondences (e.g., ⟨*ch*⟩ > ⟨*tz*⟩).

## 7. Conclusion

Cognate identification aids linguists with language comparison and reconstruction and improves the performance of machine translation systems, particularly for low-resource languages. This project examined whether a neural model incorporating character-level information could succeed at identifying cognates in a dataset of Romance and Germanic word pairs and at transferring that knowledge to identification of previously unseen Basque-Spanish cognate pairs. For each word pair in the Romance-Germanic training set, a one-hot encoding was concatenated to the output of a character-level language model trained on cognate pairs before being used to train two additional layers. After training, performance on Romance-Germanic and Basque-Spanish cognates was evaluated. Performance on the Basque dataset was evaluated when pre-trained on the Romance-Germanic data alone, fine-tuned on the Basque data, and trained on the Basque data alone.

The model was indeed able to identify the Romance-Germanic and Basque cognates. All models outperformed the baseline support vector machine (similar to that used in the most effective non-neural studies). The model performed best at identifying the Basque cognates when pre-trained on the Romance-Germanic data and fine-tuned on the Basque data; in addition, even the model trained on Romance-Germanic data alone outperformed the model trained on Basque data alone. This suggests that the model was able to extract general characteristics of cognates from the Romance-Germanic datasets and apply them to identifying cognates between Spanish and (previously unseen) Basque.

Examination of the words that the model classified correctly or incorrectly suggests that the words most difficult to classify were those with low character overlap and containing sequences of characters or character correspondences rarely seen elsewhere in the dataset. This notwithstanding, the model correctly classified words whose character correspondences resulted from a variety of phonological, orthographic, and morphological changes.

The differences in performance on the Basque-Spanish dataset between the pre-trained model and the model fine-tuned on Basque-Spanish data were also examined. The fine-tuned model appeared better able to identify cognates with character correspondences specific to the Basque-Spanish language pair. The results suggest that fine-tuning on a small dataset helps the model to correctly classify word pairs with previously unseen character correspondences resulting from new phonological, orthographic, or morphological changes (e.g., reduplication).

The model is able to learn characteristics of cognates in general from the larger dataset and apply them to a new language pair, particularly when fine-tuned with a small amount of data from the new language pair. These results suggest that efforts to identify cognates in low-resource languages are significantly aided by pre-training on larger cognate datasets from higher-resource languages.

## 7.1. Future Work

Possible extensions of this work would be to apply it to areas where cognate identification is most used, i.e., historical linguistics and machine translation. Future work could examine whether a model could correctly determine the relationships between languages based on the presence and similarity of cognates it identifies. This work could also be applied to machine translation by examining if awareness of automatically extracted cognates improves translation quality. This application would be particularly useful for common translation pairs like low-resource Basque and high-resource Spanish.

To build on the current model, one possibility could be to train the model to identify the specific types of sound changes occurring between cognates in a mother language and a daughter language. In addition, the use of phonological data could help to capture correspondences between sounds represented by different characters in different languages.

## 8. Code

The code and datasets for this project are available at `github.com/efleisig/cognate_nn_iw`. The models were trained and tested on the Princeton Adroit computing cluster.

## 9. Acknowledgments

Many thanks to Professor Christiane Fellbaum for her invaluable feedback and support throughout the semester, which made completing this independent work a wonderful experience. Special thanks as well to Professors Danqi Chen and Karthik Narasimhan for their guidance on character-level models. As always, thank you to the Princeton Research Computing group for their support with the Adroit cluster.

## 10. Honor Code

This paper represents my own work in accordance with University regulations.

Eve Fleisig

# References

[1] Q. D. Atkinson, "The descent of words," *Proceedings of the National Academy of Sciences*, vol. 110, no. 11, pp. 4159–4160, 2013. [Online]. Available: https://www.pnas.org/content/110/11/4159

[2] L. Beinborn, T. Zesch, and I. Gurevych, "Cognate production using character-based machine translation," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 883–891.

[3] V. Chikin and I. Ilin, "Chars2vec: character-based language model for handling real world texts with spelling errors," 2019.

[4] G. de Melo, "Etymological Wordnet: Tracing the history of words," in *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, Eds. Paris, France: European Language Resources Association (ELRA), May 2014, pp. 1148–1154. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1083_Paper.pdf

[5] G. Jäger, J.-M. List, and P. Sofroniev, "Using support vector machines and state-of-the-art algorithms for phonetic alignment to identify cognates in multi-lingual wordlists," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1205–1216. [Online]. Available: https://www.aclweb.org/anthology/E17-1113

[6] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," *Machine Translation*, vol. 32, no. 1-2, pp. 167–189, 2018.

[7] D. Krajewska and T. H. Godoy, "Handling reduplication in basque: a problem for spell checking," *Procesamiento del lenguaje natural*, vol. 47, pp. 277–281, 2011.

[8] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet physics doklady*, vol. 10, no. 8, 1966, pp. 707–710.

[9] J.-M. List, "Lexstat: Automatic detection of cognates in multilingual wordlists," in *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, 2012, pp. 117–125.

[10] S. Malmasi and M. Dras, "Cognate identification using machine translation," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, 2015, pp. 138–141.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[12] M. Morvan, "Dictionnaire étymologique basque en français-espagnol-anglais [Basque etymological dictionary in French-Spanish-English]," 2016. [Online]. Available: http://projetbabel.org/basque

[13] N. Pourdamghani and K. Knight, "Neighbors helping the poor: improving low-resource machine translation using related languages," *Machine Translation*, vol. 33, no. 3, pp. 239–258, 2019.

[14] T. Rama, "Automatic cognate identification with gap-weighted string subsequences." in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1227–1231.

[15] T. Rama, J. Wahle, P. Sofroniev, and G. Jäger, "Fast and unsupervised methods for multilingual cognate clustering," *arXiv preprint arXiv:1702.04938*, 2017.

[16] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," *CoRR*, vol. abs/1508.07909, 2015. [Online]. Available: http://arxiv.org/abs/1508.07909

[17] E. Soisalon-Soininen and M. Granroth-Wilding, "Transfer learning for cognate identification in low-resource languages," in *First Workshop on Typology for Polyglot NLP*, 2019.

[18] M. Swadesh, "Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos," *Proceedings of the American philosophical society*, vol. 96, no. 4, pp. 452–463, 1952.

[19] I. J. Unanue, L. G. Arratibel, E. Z. Borzeshi, and M. Piccardi, "English-basque statistical and neural machine translation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[20] A. Üstün, M. Kurfalı, and B. Can, "Characters or morphemes: How to represent words?" in *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018, pp. 144–153.

[21] J. L. White and M. Horst, "Cognate awareness-raising in late childhood: teachable and useful," *Language Awareness*, vol. 21, no. 1-2, pp. 181–196, 2012.