IMPROVING NEURAL LANGUAGE MODELS ON LOW-RESOURCE CREOLE LANGUAGES

BY

SARAH SCHIEFERSTEIN

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Adviser:

Associate Professor Julia Hockenmaier

ABSTRACT

When using neural models for NLP tasks, like language modelling, it is difficult to utilize a language with little data, also known as a low-resource language. Creole languages are frequently low-resource and as such it is difficult to train neural language models for them well. Creole languages are a special type of language that is widely thought of as having multiple parents and thus receiving a mix of evolutionary traits from all of them. One of a creole language's parents is known as the lexifier, which gives the creole its lexicon, and the other parents are known as substrates, which possibly are thought to give the creole language its morphology and syntax. Creole languages are most lexically similar to their lexifier and most syntactically similar to otherwise unrelated creole languages. High lexical similarity to the lexifier is unsurprising because by definition lexifiers provide a creole's lexicon, but high syntactic similarity to the other unrelated creole languages is not obvious and is explored in detail. We can use this information about creole languages' unique genesis and typology to decrease the perplexity of neural language models on low-resource creole languages. We discovered that syntactically similar languages (especially other creole languages) can successfully transfer learned features during pretraining from a high-resource language to a low-resource creole language through a method called neural stacking. A method that normalized the vocabulary of a creole language to its lexifier also lowered perplexities of creole-language neural models.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

Neural networks have been the mostly-uncontested model of choice for natural language processing research over the last five years [1], as of 2018. With their innate ability to describe complex functions on what is usually millions of parameters, they yield excellent results when there is enough data to train the models. Unfortunately, due to the large number of parameters in neural networks, languages that do not have large amounts of data do not perform as well on these models; as such, there is much research devoted to what I will refer to as *low-resource methods*. While lack of data is an apparent issue for neural networks, other machine learning methods also use these methods to enhance their performance. Languages with not much data are *low-resource*, and through either augmenting the data artificially [2], projecting the unlabelled data onto labelled data [3, 4], creating special models [5, 6, 7], or being clever with how the data is used in the model, greatly improved results are achieved.

Virtually all creole languages are low-resource languages. When machine learning models are used to process them, they benefit from low-resource methods. Notably, there are a few machine learning models that use low-resource methods that take advantage of 'unique' aspects of creole languages [2, 8, 9]. When I say 'unique', I refer to the fact that while most languages evolve from one parent, creoles evolve from several parents into one 'mixed' language. For example, in the case of Haitian Creole, its parents are French and several West African languages, one of which is assumed to be the language Fon. French contributes to the language's lexemes, while Fon and other African languages contribute to the morphosyntax of the language. As such, French would be Haitian Creole's *lexifier* or *superstrate*, and Fon would be one of Haitian Creole's *substrates*. Note that while virtually all languages borrow linguistic features from other languages (usually vocabulary), they still only have one evolutionary parent.

The goal of this thesis is to create low-resource methods that will increase the performance of neural natural language processing tasks, in particular neural language modelling, when the target language of these tasks is a creole language. Language modelling is the task I focus on because there is a higher amount of creole language data available for the task, but the methods proposed in this thesis can apply to other tasks and models as well. When the unique genesis of creole languages is studied, along with the interesting typology of creoles as they exist today, it is natural to think that similarly unique methods could be created to process this group of languages. However, this is confounded by another innate fact of neural networks; it is very difficult to understand why certain models work better

than others. Tweaks to a neural model that make sense by linguistic hypotheses (e.g. creole genesis theory and typology) could have no basis in the neural model's empirical score (e.g. the perplexity of a language model). This is made worse by the fact that each experiment will necessarily use a target language with very little data; when there is very little data, the neural network performs rather poorly already because it does not have nearly enough data to converge to good parameters, so it is possible that no low-resource trick could push the model closer to convergence and to a better score. A low-resource method could even make the model perform worse by this logic, no matter how correct the linguistic theory behind the method is.

Chapter 2 will review the history and linguistic theory of creole languages and their genesis. Chapter 3 will review existing research that focuses on approaching creoles computationally in unique ways. Chapter 4 will examine how phylogenetic estimation methods that do assume a single parent per language behave with creole languages, which have multiple parents. Chapter 5 will describe the neural language model that will be used as a baseline for further experiments. Chapter 6 will describe the low-resource methods that will augment the base neural language model. Chapter 7 describes the data that will be used to test the neural models created. Chapter 8 tests the low-resource methods and examines the relative changes in perplexity per target creole language to evaluate the efficacy of each model.

CHAPTER 2: REVIEW OF CREOLE LANGUAGES AND THEIR GENESIS

Creole languages have been known to exist for centuries, but only recently have they begun to be seriously researched by linguists. In the past, they were regarded as either inferior versions of their lexifiers, or not natural languages and thus not worth studying. Both of these views have been disregarded recently, and study into creoles, and particularly their genesis, has become hotly debated. Still yet, there does not exist a consensus on what linguistic processes created these languages. This chapter attempts to provide an overview on the many theories of creole genesis that scholars have published over the last century.

2.1 REVIEW OF CREOLE LANGUAGES

First, I will begin with a review of creole languages. The definitions of creoles and pidgins are not entirely agreed upon, but one view is presented here. The majority of scholars [10] agree on the usual sociohistorical background that led to the creation of creoles. This background is one in which enslaved humans who spoke different languages were transported to a place where a (usually European) language was spoken. (Note that this is not the only sociohistorical background there are different scenarios, such as those that created the creole languages Hawaiian Creole English and Sango). The enslaved humans had a need to communicate with each other, so they used elements of the European language to create a language for everyday communication between themselves [11]. The European language provides the lexicon of the language and is hence named the lexifier. This language is generally thought to be more regular and simpler, in terms of how easy it is to describe the rules of the language (i.e. less consonant clusters, less phonemes, less pronoun differentiation, less lexicon, less morphology) [11]. Most people call this a pidgin; eventually, the children that learn this language natively are said to creolize it, thus making the language a creole; from then on, the language develops complexity and irregularity like any other language. Note that the morphosyntax and structure of creoles are generally thought to be similar to each other, regardless of their historical context [11]. Beyond this accepted background, the question then becomes one of creole genesis: what linguistic processes actually cause the creation of the creole?

2.2 MONOGENESIS

One theory of creole genesis is monogenesis. Monogenesis states that all pidgins and creoles come from the same language, namely West African Pidgin Portugeuese [11, 10]. This pidgin is said to have been spoken on slave trade ships for several centuries. The monogenesis hypothesis posits that all pidgins and creoles have simply been relexified versions of this language; that is, the Portuguese-derived lexicon of the pidgin was replaced by words of another lexifier. Since through the process of relexification the structure of the language stays the same, this is how this theory explains that all creoles are similar in structure to one another. An obvious critique of this hypothesis is that many creole languages are not based on European languages, so they could not have been relexified from the Portugeuese pidgin due to geographical location.

2.3 SUBSTRATUM THEORY

Another theory of creole genesis is the substratum theory, which posits that the substrates involved in creating a creole supply the structural features of the language. The substrates of a creole are the non-dominant languages; in the background case previously described, the substrates would be the different African languages the slaves spoke. One version of this theory is also called relexification, but instead of relexifying the lexicon of West African Pidgin, the lexicon of the substrates are relabeled based on the dominant languages lexicon [10]. As such, this hypothesis claims that a creoles structure is provided by some mixture of the substrates structure. The lexicon is still provided by the lexifier by the relexification process. Another version of this theory is called code switching [12]. Code switching is already a concept in bilingualism where a person who knows two languages either mixes or switches between the two while speaking. This type of code switching is slightly different than the bilingual version. The hypothesis states that the slaves trying to speak the lexifier language did not have the resources to fully learn it, so they used the morphosyntax of their own languages with the lexicon of the lexifier to approximately speak it, thus creating a creole. Once again, this code switching hypothesis claims that the substrates provide the grammar and structure of the creole.

2.4 SUPERSTRATUM THEORY

Yet another theory is the superstratum theory. This theory [13] does not necessitate that a creole comes from a pidgin. Instead, going back to the original creole background,

it arises when non-European slaves interact with European slaves/indentured servants (who usually speak non-standard dialects of the European language), so the creole language that the slaves communicate with is just this non-standard dialect of the lexifier, or superstrate. As such, this would make creoles a direct descendant of the European language instead of a mix of several languages. Note that this theory does not account for the similarity across all creoles (this is because the authors of such theories [12, 13] disagree with the fact that all creoles share similar structure; they argue that the class of languages called creoles are not structurally similar, but simply share a similar social history).

2.5 UNIVERSALIST THEORY

The last main theory is the universalist theory. One version of this theory is called the language bioprogram hypothesis [11, 12]. Instead of claiming that creoles structural similarity comes from their component languages, this hypothesis claims that the similar structural features come from when the language is nativized by a child. It states that when the child learns a pidgin, since the pidgin is unstructured and gives the child a poverty of linguistic input [12], the child is forced to create parameters in the pidgin it is learning because otherwise the parameters are unmarked. As such, the Universal Grammar steps in to fill these parameters in a way that is biologically inherent to all members of the species; since all creoles are learned by children from very simple pidgin languages, all creoles must exhibit similar structural features. In this thesis, that is how they explain the structural similarity between creoles. Adjacent to this theory is an effort to prove that creoles are typologically similar in structure [14]. Some articles use the WALS features [15] from APiCS [16] (a database of features of creoles and pidgins) to describe their similarity [17]; others use techniques from computational biology to show that they are related [18]; others simply compare large amounts of creole languages to prove that they have similar features [19]. Critiques against these methods and the universalist theory include that this finding comes from a generalization, and many counterexamples can be provided [11, 12]. Another critique is that people who classify creoles similarly may be guilty of a circular definition of creoles, wherein creoles that do not fit the typological classification they made for creoles are deemed not creoles, instead of defining something as a creole purely by its sociohistorical context.

2.6 OTHER CREOLE GENESIS THEORIES

Many scholars also make room for mixes of the above theories. For example, one article [20] argues that the substrate and universalist hypothesis complement each other. Where the substrate hypothesis cannot explain a structural feature, perhaps the poverty of the pidgin language that spawned the creole created the structural feature universally. Likewise, where the universality hypothesis cannot explain a structural feature, like when a feature is very unlikely to have come about universally, one of the substrates may contain that structural feature and explain its existence. Another work [12] identifies that since many creoles generate under different circumstances, it is possible that some creoles exhibit tenets of certain hypotheses more than other creoles due to this difference. For example, the author of this work writes that while the superstrate hypothesis does not seem likely for most creoles, the hypothesis seems to be correct in the case of one particular creole, Reunionnais Creole, because of its sociohistorical context and the fact that it is very close to its lexifier, 18th century French, in structure.

2.7 THOUGHTS AND TYING LINGUSTIC THEORY TO COMPUTATION

As a final aside, I will provide my personal opinion on the various genesis theories in the creolistics community. It seems to me that many of the arguments I have mentioned were written before this community had access to the APiCS database in 2013, which provides the features of 76 pidgins and creoles, as well as larger corpora. Because of this, it was common to disrepute arguments based on pure reasoning or narrowly focusing only on specific examples to prove a point while ignoring broad patterns in linguistic data. One example of such a point made in this manner instead of empirically is one article [12] arguing against creoles sharing a typological class because they were made in different sociohistorical contexts, despite the fact that other sources support the idae that creoles generated across different contexts and geographies still fit into a similar typological class [17, 18]. Essentially, such arguments have not aged well in this field as the amount of relevent data increases.

I hope in the future we can come closer to a regular study of creole genesis, as opposed to the current climate of many combatting viewpoints. Perhaps this can be accomplished by making sure to consider the diachronic history of how each individual creole was generated (as emphasized in [12]), and allowing room for a mix of each potential hypothesis to explain the aspects of any particular creoles genesis.

All of this information influences what types of low-resource methods should be used with creoles. This will be explained in detail in chapters 3 and 6. One obvious inference

from this information is that creoles have a very similar lexicon to their lexifier. Lexifiers are often the language of colonizers from Europe, so lexifiers are *high resource* and have plenty of data. Even if a creole is extremely low-resource, at the very least an approximation of what each word 'means' by itself (but not necessarily as it relates to other words nearby it) can frequently be derived from words that sound or look alike in its lexifier; this is information that can help a model with a low-resource creole target.

CHAPTER 3: REVIEW OF COMPUTATIONAL PROCESSING ON CREOLE LANGUAGES

This section presents a review of two types of papers: one that seeks to discover properties of creole genesis through computation, and another that uses the typology of creoles and creole genesis ideas to aid natural language processing tasks on creole languages.

3.1 COMPUTATION FOR CREOLE GENESIS

3.1.1 Explorations in creole research with phylogenetic tools

This paper by Daval-Markussen and Bakker [18] studies how to research the structure of creole languages and how a creole language compares with its superstrate and substrate languages. It does this by using phylogenetic tools, like Splitstree, to create networks that explain and compare the languages' structure based on an assumed model of character evolution. It focuses on the structure because creole structure evolves from multiple parents, while the lexicon usually evolves from one parent (the superstrate) and is less interesting and controversial to study. The paper seeks to prove that creole research can contribute greatly to the study of language evolution and genesis by the fact that they are complexly and (possibly) regularly generated by multiple parents as opposed to the usual single parent in the majority of languages. This paper focuses on English-based creoles (note that this is a questionable constraint; why not explore if creoles based on other lexifiers are typologically different in structure?)

The first task put the data combined with geographical categories into SplitsTree to output a visual network (which defines a tree initially, but becomes a network to represent uncertainty in the phylogeny estimation) describing the evolution of these languages. They organized data for 33 Atlantic Creoles; each has 62 features denoting the absence or presence of certain phenomena in the languages. In the output network, genealogical clusters exist that affirm previous research on the respective languages' heredity, as well as clusters that represent geographical closeness.

The second task focuses on using a tool to research the possibilities of various existing creole genesis theories, which we will briefly redefine here. They tested the superstratist view which states that creole structure comes from the dominant language, the substratist view which states that creole structure comes from the substrate languages, the feature pool view which states that creole generation comes from an evolutionary-esque battle for survival amongst its composing languages' features, and the universalist view which posits that creole

structures are similar and were all borne out of the challenge of communicating across several lanaguages imperfectly and thus developed universally in the same way to conquer this challenge. For each view, they output a network which tested the views (e.g. putting in substrate and/or superstrate non-creoles and respective creoles) and loosely examining how they 'grouped' in the network. In each network, the creoles grouped together closely and the non-creoles were rarely near the creoles; as such, the networks support the universalist view.

The third task seeks to research the typology of creole languages. Specifically, it seeks to discover if creoles, as a group, are structurally distinguishable from non-creoles. It does this using a Neighbor-Joining trees visualization, which can show groupings (genesis of languages is not relevant in this task). It uses data consisting of creoles and non-creoles, and morphosyntatic features that say 'yes' or 'no' to presence of a feature; once again, the creoles group together and the other languages are not near them. From this point, several different datasets with different combinations of creoles and non-creoles, as well as differing and more complex features are used and all yield the same result of creoles clustering; the typological distinction of creoles increases when more languages are included in the data.

3.1.2 Statistical Modeling of Creole Genesis

This paper by Murawaki [17] seeks to challenge the tree model used to represent the evolutionary history of languages, like in [18] described above. This paper uses creole languages as an extreme example to challenge a tree-like model of evolution notion because they notably have more than one parent. To accomplish this, data-driven methods are used.

Crucially, this paper makes an extremely salient observation on [18]'s usage of phylogenetic tools: it exclusively used tools that assume that each language evolves with a single parent, even though the paper attempts to study creole languages which evolve with multiple parents. This fact makes the phylogenetic analysis presented as-is less powerful. Also importantly, it questions the paper's validity in claiming that examining typology can support any creole genesis theory, given that genesis happened in the past, and studying typology only considers modern language features.

The first task of this paper is to examine creole languages' distinctiveness not using tree-based model. They used similar data to [18], which is several features denoting the absence or presence of certain phenomena in the languages. They attempt to use a linear SVM classifier to categorize languages as creole or not in order to test creole distinctiveness. The SVM failed to perfectly categorize creoles and non-creoles given this feature data, despite previous work hypothesizing that they could be perfectly separated based on typo-

logical features. Next, PCA was executed on the data. Two clusters of non-creole languages were distinct from the one creole language cluster. Despite this clustering, there were some exceptions; some non-creole languages were found near the creole cluster, and some creole languages were found closer to non-creole cluster. While the clusters are distinct, the creole and non-creole languages notably overlap in the PCA visualization.

The second task of this paper is to create a model intended to be able to generate creole languages. It assumes that a creole, defined as a set of features, can be stochastically generated by mixing the features of a lexifier, substrates, and 'something else'. The model infers these mixing proportions from observed data, which is based on ancestral mixture concepts from computational biology. The simulated creole languages from this data mostly picks features from the 'something else' category as opposed to the lexifier or substrate, which supports the universalist creole theory and puts doubt on a feature pool theory (which is described in chapter 2).

3.2 CREOLE THEORY FOR NLP

3.2.1 CMU Haitian Creole-English Translation System for WMT 2011

This paper by Hewavitharana et al. [2] describes a Haitian Creole to English statistical machine translation system. Its main goal was to create a system that would translate SMS messages well by using texts written in Haitian Creole sent during the Haiti earthquake. I will focus on the method they used that relates directly to unique aspects of creole languages.

The dataset was composed of data given to them in the competition that included SMS data and other parallel Haitian Creole/English corpora. Some of the SMS data was already manually cleaned, but some of it was uncleaned and difficult to translate as-is. As such, the raw SMS data received special preprocessing. Importantly, this included collapsing OOV words that were obviously mistakes (had special characters in them) into the nearest vocabulary words as determined by shortest edit distance to the clean SMS data's lexicon, as well as normalizing spelling. Note that Haitian Creole tends to have variation in its orthography, leading to many of these alternate spellings of identical words, making this process important. The spelling normalization was done using French because Haitian Creole did not have enough data to execute this normally using the noisy channel model. They did this based on the knowledge that Haitian Creole's lexicon derives from French, so they ignored the words in the English vocabulary (because the texts contained those as well) and normalized all other words according to the French probability distribution and into French

words with French orthography.

3.2.2 Anou Tradir: Experiences In Building Statistical Machine Translation Systems For Mauritian Languages Creole, English, French

This paper by Dabre et al. [8] describes experimenting with statistical machine translation systems involving Mauritian Creole, English, and French. Particularly, it experiments with using French as a bridge language for translating English into Creole via the 'transfer method'.

It begins by describing and comparing the grammar and lexicon of Mauritian Creole to English and French. Essentially, the paper claims that while the Creole has some English influences (because the nation was ruled by English speakers for a while), it is influenced by French significantly more in both lexicon and grammar (essentially, it claims that French is Mauritian Creole's lexifier).

They developed SMT systems for all combinations of English, French, and Creole translations that included Creole, and additionally included the system where English is translated into Creole by a French bridge. The English-French corpus had 80x more lines than the manually created corpora for Creole-French and Creole-English.

They used a previously proposed 'transfer method' to get to English to Creole by using French by translating English-to-French, then French-to-Creole. Both translation steps must be good for this method to yield usable results. This goes off the assumption that English-French is good, and that French-Creole is good despite the small corpus size because the languages are close. The exact method is in the paper and involves scoring several sentences and intermediate translations of these sentences.

They applied each translation system to a 'hard' corpus (longer sentences) and an 'easy' corpus (shorter sentences). For easy sentences, the English-to-Creole bridge translation worked better than the direct. For hard sentences, the direct translation worked better than bridge translations. They posit that this is because the French to English hard translations did not score well, so the bridge translation degraded multiplicatively (as said before, in order for this method to work well both translations must perform well).

3.2.3 Universal Dependencies Parsing for Colloquial Singaporean English

This paper by Wang et al. [9] seeks to create a Singlish treebank for parsing. The existing Singlish treebanks are small, but the paper hypothesizes since that English has lexical and syntactic similarities to English and a large treebank, it can augment the small Singlish

treebank to create Singlish dependency trees with a neural dependency parsing model using biaffine attention (using neural stacking to integrate the English portion, which is explained in chapter 6 in detail). A neurally stacked POS tagger is trained similarly because POS tags are used in dependency parsing. They note that there are many influences from languages that are not English, and notes that mixing English into the training cannot account for many syntactical and lexical occurrences in Singlish.

A POS tagger for Singporean English is created via neural stacking on English. Briefly, this means that an English POS tagger was trained regularly. Afterward, a conjoined model consisting of the pretrained English POS tagger as the first layers and the untrained Singaporean POS tagger as the last layers; the input goes through the English tagger first, and then its emission vector (i.e. the final result of the neural network before it is translated into a part of speech category) is concatenated with the input embedding of the Singaporean POS tagger, which feeds forward as normal. This conjoined model is trained and loss is backpropagated through it entirely. Essentially, the features learned by the English tagger are transferable to the Singaporean tagger via this method, which allows the Singaporean model to converge further without having more data in its own language. A model for dependency parsing is generated in a very similar manner.

For both POS tagging and dependency parsing, the stacked model performs better than the unstacked baseline models (English alone or Singlish alone) according to relative error reduction for POS and UAS and LAS ratings for dependency parsing.

3.3 CONCLUSIONS

The first paper [18] shows that phylogenetic methods that try to reconstruct language evolution based on syntactic features places creoles closer in the reconstruction rather than lexifiers and substrates, showing that creoles are similar syntactically. The second paper [17] is important because it shows that while creoles are not linearly separable from non-creoles, they are distinctive. Its explorations into creating non-tree, mixed-parent models are fascinating as well and possibly support universalist theories of creole genesis. The third paper [2] introduces the fascinating idea of normalizing the spelling of creole languages effectively with the very lexically similar lexifier language. The fourth paper [8] is important because it acknowledges that the lexifier has both lexical and syntactical similarities to the creole language, which makes it useful to use in a unique transfer learning machine translation method. The fifth paper [9] suggests a neural model which yields excellent results on a low-resource creole model when used with a similar high-resource language.

Very interestingly, all 3 machine learning papers utilized the lexifier, and some made

claims that the lexifier was the best choice of language to use in their models due to its close syntax to a creole language. While lexifiers do tend to have a slightly closer syntax to creoles compared to an arbitrary language, as evidenced by the data-driven explorations in the first two papers, substrates and especially creole languages have much closer syntax. Consider [9]: its part of speech tagging and dependency parsing models' success depend directly on how well the models are able to capture the syntax of a language. The lexicon itself is abstracted away by the input layer embeddings used (they are turned into feature vectors representing a word's semantics, usually independent from explicit lexicons), which makes the choice of English seem silly; why justify the use of English based on syntax when other languages, like other creoles, have much closer syntax to Singaporean English? More broadly, this line of thought makes clear a certain unanswered but very important question: in these last 2 machine learning models we reviewed, how important is the syntactic similarity of the language that aids the low-resource language? The claim that having similar syntax is the reason for the choice of lexifiers aiding low-resource creole languages is made, but is not examined further. This question will be researched later using the neural stacking model with languages of varying syntactic closeness.

CHAPTER 4: STANDARD SINGLE-PARENT PHYLOGENY ESTIMATION ON CREOLE LANGUAGES

As an extension to the type of phylogeny estimation exploration performed in paper [18] described in detail in chapter 3, we will examine how RAxML [21], a maximum likelihood phylogeny estimation software, places creole languages in its estimated trees. The placement will always be 'incorrect' because of RAxML's single parent assumption in its evolution models, but their location in the tree will describe the creole languages' syntactic typology relative to other languages. There exist papers that already include creole languages in their language evolution 'tree of life', like [22]. This paper and others use lexicostatistical methods that group any pair of languages closer together if they share identical cognates. Unsurprisingly, since creoles take the lexicon of their lexifier, the creoles end up very close to the lexifier in these trees. This is why in this chapter we will study syntactic typology; it is more exciting to explore. The input to RAxML uses Indo-European languages except for two creole langauges, Sranan and Haitian Creole. Sranan and Haitian Creole's lexifiers, English and French, are in the dataset as well.

4.1 DATA

Morphosyntactic features from APiCS [16] that correspond with WALS [15] features were downloaded. The features used and their corresponding APiCS number are displayed in Table 4.1. From this data, we can represent the WALS features as a matrix. Each row represents a language, and each column represents a morphosyntactic feature from WALS/APiCS. Each entry represents the particular language's value for the corresponding feature as an integer. Missing data is imputed with nearest neighbor imputations via python library fancyimpute. The data in PHYLIP format can be viewed here: `https://github.com/aviolaine/nlp_for_creoles/tree/master/raxml`. Both the categorical matrix and the binary-encoded matrix can be viewed. The features that correspond to the categorical matrix's columns are also shown there.

4.2 EVALUATION METHOD

As inspired by [23], I will evaluate the sections of the tree that do not contain creole languages by comparing the tree to the generally approved ground truth. Since I am only using Indo-European languages, the ground truth is very accurate. To be specific, for each

APiCS Feature Number	Feature Name
13	Gender distinctions in personal pronouns
7	Order of relative clause and noun
29	Indefinite articles
58	Alignment of case marking of full noun phrases
73	Predicative noun phrases
103	Polar questions
100	Negative morpheme types
59	Alignment of case marking of personal pronouns
2	Order of possessor and possessum
33	Distance contrasts in demonstratives
71	Noun phrase conjunction and comitative
76	Predicative noun phrases and predicative locative phrases
18	Politeness distinctions in second-person pronouns
72	Nominal and verbal conjunction
4	Order of adposition and noun phrase
54	Suppletion according to tense and aspect
70	Comitatives and instrumentals
36	Sortal numeral classifiers
6	Order of cardinal numeral and noun
8	Order of degree word and adjective
56	The prohibitive
60	Ditransitive constructions with give
91	Applicative constructions
23	Expression of nominal plural meaning
92	Subject relative clauses
62	Expression of pronominal subjects
32	Pronominal and adnominal demonstratives
77	Predicative possession
22	Occurrence of nominal plural markers
1	Order of subject, object, and verb
12	Position of interrogative phrases in content questions
38	Marking of possessor noun phrases
42	Comparative standard marking
3	Order of adjective and noun
15	Inclusive/exclusive distinct. in indpt. personal pronouns
21	Indefinite pronouns
28	Definite articles
5	Order of demonstrative and noun

Table 4.1: APiCS/WALS features used in phylogeny estimation

Indo-European subfamily, if the subfamily is not exactly a clade in the estimated tree, I will note how the tree deviates from the subfamily.

4.3 METHODS

Maximum likelihood is used to estimate a phylogeny estimation with RAxML. The input data is transformed into binary features instead of categorical via a one-hot encoding (this means that each entry stands for the absence or presence of a feature value). The default settings are used at `http://www.trex.uqam.ca/index.php?action=raxml`, but data type is set to BINARY and substitution model is set to BINCAT, such that the phylogeny is estimated based on a model that assumes that each binary character can switch from present to absent and vice versa over time as languages evolve. Specifically, RAxML version 7.2.6 as released by Alexandros Stamatakis in February 2010 [21] was used with the arguments "./raxmlHPC -m BINCAT -f d -s inputRaxML.phy -n trex". Note that RAxML does not produce deterministic results, so the results shown here won't necessarily be exactly reproduced with identical input and commands, particularly branches with low support. This is why the four trees that are produced slightly differ in estimating Indo-European, non-creole languages; they decided to put the Indo-European languages with low confidence in different places each time. RAxML is chosen because it is a standard, well-performing method that has a single-parent evolutionary assumption. As such, we can examine how creole genesis is explained under this likely false evolutionary model.

4.4 EVALUATION OF RESULTS

See the generated trees in Figure 4.1-4.4. These figures can also be seen here: `https://github.com/aviolaine/nlp_for_creoles/tree/master/raxml`. Each language name is limited to 5 characters maximum in the figure (e.g. Sranan is cut short to Srana). The languages used, not including the creoles, are from the Indo-European subfamilies Albanian, Baltic, Celtic, Germanic, Indic, Iranian, Romance, and Slavic. I define their subfamilies identically to how the WALS database defines them. Some subfamilies have more languages than others in these experiments. Four trees are estimated: one with all of these languages, one with all of these languages except Sranan, one with all of these languages except Haitian Creole, and one with all of these languages without the creoles.

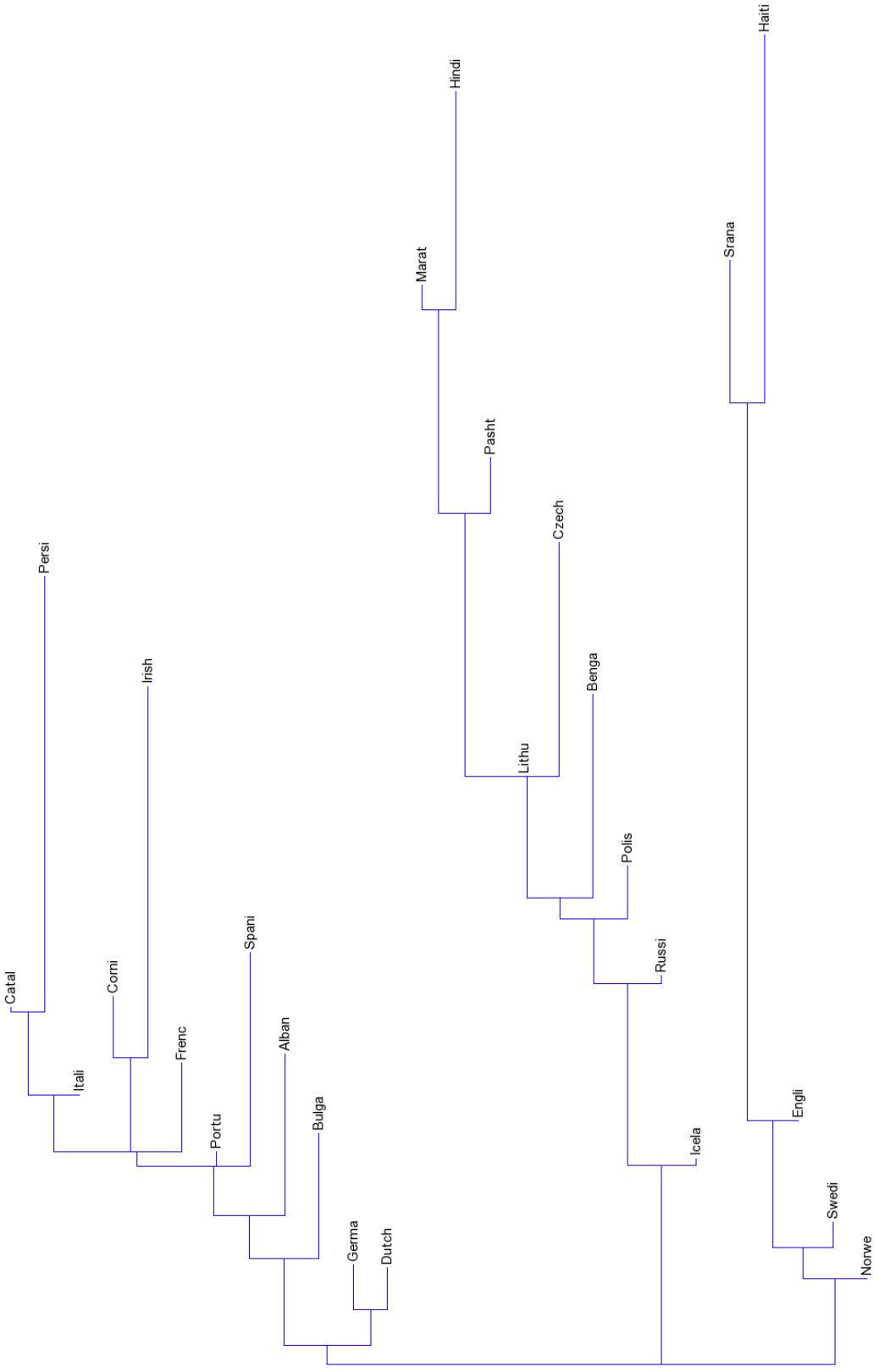


Figure 4.1: Both creoles

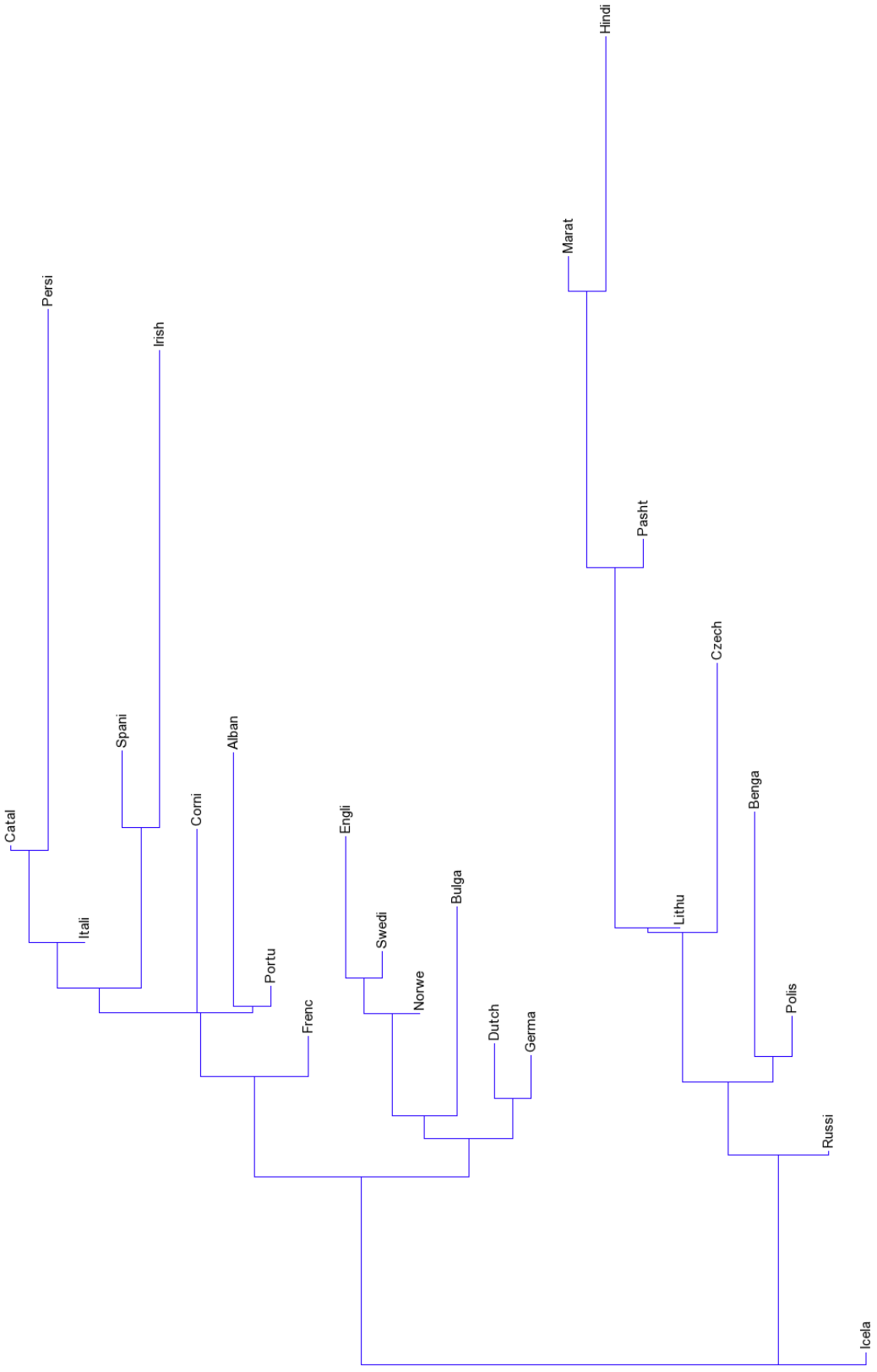


Figure 4.2: Neither creoles

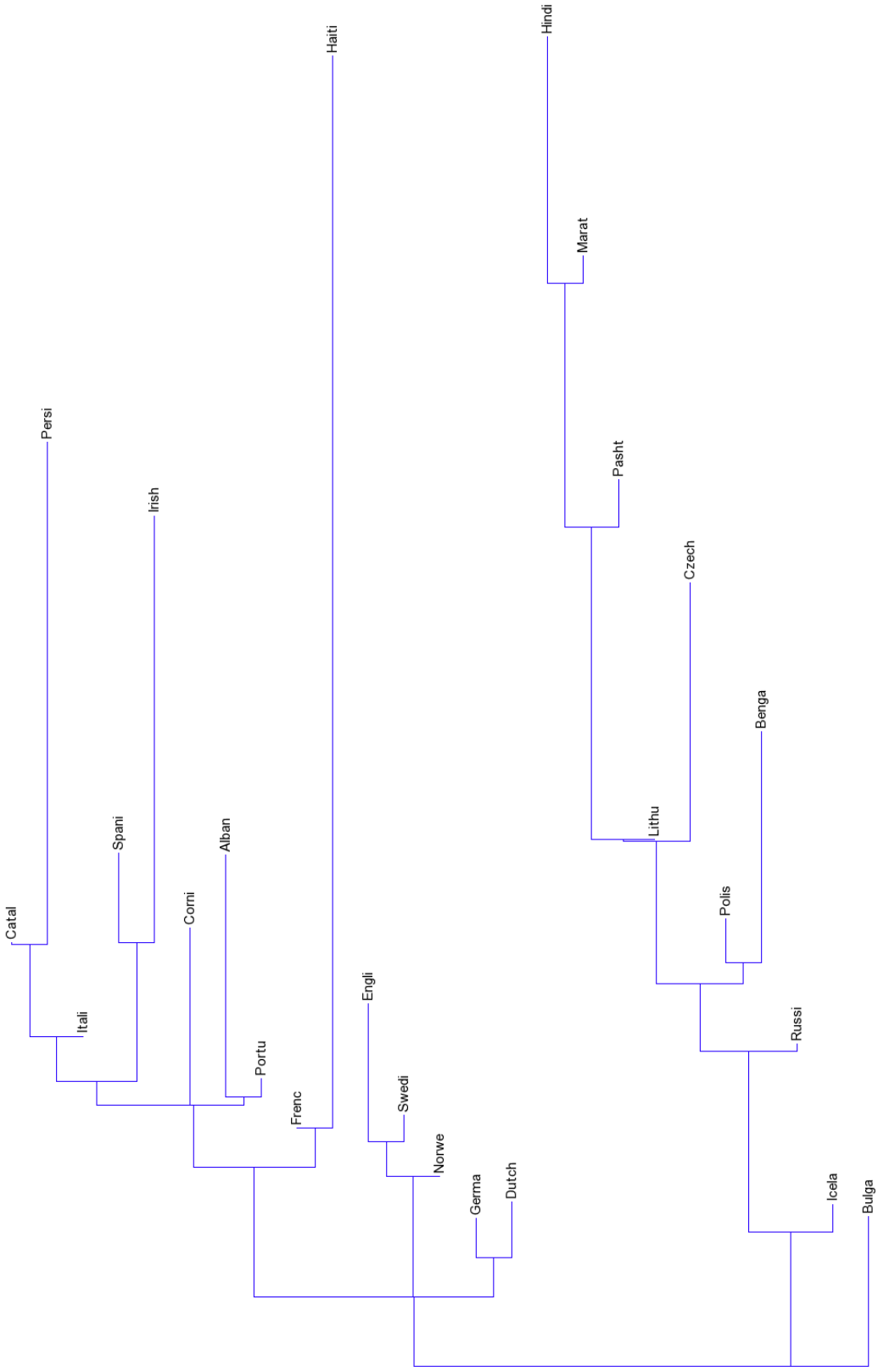


Figure 4.3: Only Haitian Creole

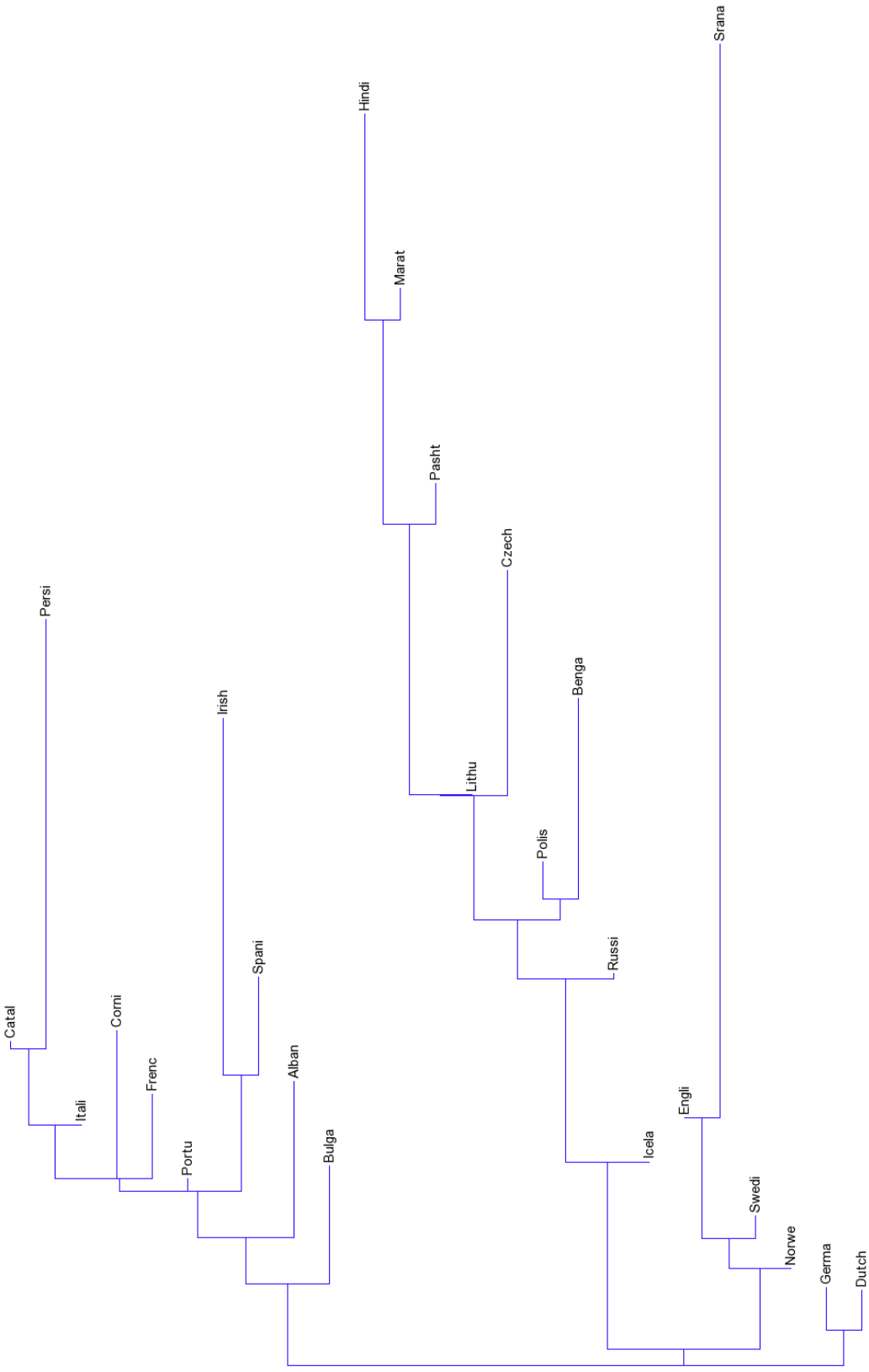


Figure 4.4: Only Sranan

- **No creoles:**
  - Celtic: Irish and Cornish are placed in the Romance family, but are closer than not.
  - Germanic: All of the Germanic languages are close.
  - Indic: Bengali is separated from the other Indic languages by Pashto and Balto-Slavic languages.
  - Iranian: Pashto and Persian are extremely far apart.
  - Romance: All of the Romance languages are close.
  - Slavic: Bulgarian has grouped with Germanic languages.
- **Just Sranan:** Sranan appears nearest to English in the Germanic section, but has a large distance from English. English is Sranan's lexifier, so this supports a creole having a similar lexifier, at least compared to many other arbitrary languages, in terms of syntactic closeness.
- **Just Haitian Creole:** Haitian Creole is sibling to French and has a long branch distance to French. This also supports that a creole has a syntactic similarity to its lexifier.
- **Both creoles:** Haitian Creole and Sranan are siblings. The sibling of their clade is English, which is still a large branch distance away. Haitian Creole is farther from English than Sranan. While English is close to Sranan, it is even closer to another completely unrelated creole language, which supports the universal theory of creole genesis; note that this finding doesn't disregard the fact that lexifiers can be syntactically similar to creoles as well. This is an interesting result, because Sranan and Haitian Creole developed in two separate countries, and their existence in the same region cannot alone account for their syntactic closeness compared to all other Indo-European languages in this data.

4.5 SUMMARY OF FINDINGS

Even though RAxML's binary model of evolution on syntactic WALS feature is inappropriate because the model of evolution is tree-like, the trees produced still say much about the morphosyntactic closeness of creole languages compared to themselves and other languages. We can see that according to RAxML, the two creole languages studied are

syntactically close to their lexifier when compared to any arbitrary language, but the two creole languages are *much* closer to each other than they are to their lexifiers. Lexifiers indeed have similar syntactic features compared to their creole languages and are not an awful choice in the low-resource multlingual models with creole languages as a target discussed in Chapter 3, but creole languages are much closer to each other syntactically and have the potential to transfer more syntactic information. This idea should not be disregarded because the potential of transfering more syntactic information could very much boost the performance of low-resource languages on NLP tasks. Lastly, it is concerning how the RAxML trees sometimes fail at correctly estimating subfamilies based on syntactic WALS features; it's not abysmal, but it could be better and lexicostatistical trees like in [22] tend reconstruct subfamilies more accurately. Perhaps the trees found in [18] could be examined for their correctness to ground truth subfamilies as well?

CHAPTER 5: NEURAL LANGUAGE MODEL

In order to demonstrate techniques for low-resource creole languages in neural models, we will be focusing on neural language models. In particular, the baseline model is a word-level LSTM language model with pre-trained FastText embeddings used at the input layer. Language models require no labelled data, so creole languages with no labelled data can still create language models. Language models are useful in NLP because they accurately represent the distribution of words over a language given a context. In particular, having a good language model is essential for NLP tasks like machine translation [24], or any task that requires language generation. For example, some language models can easily generate language word-by-word: given a word and its context, it can choose a probable word to occur next. Also, keep in mind that the methods recommended in chapter 6 can apply to many neural models for NLP, not just language modelling.

First I will describe the language model created by Mikolov in [25], called the recurrent neural network language model or RNNLM. Its purpose is to learn patterns of variable length in sequences by mimicking how neurons in a human brain have short-term memory. Its input layer describes a single word via a 1 of N encoding, $\mathbf{w}(t)$, where a vector of length N is described where all entries are 0 except for one entry, which is set to 1, which thus identifies the vector as corresponding to a unique vocabulary word. No two unique vocabulary words can share the same vector via this encoding. This input layer is concatenated with the hidden layer from the previous timestep, which is then multiplied by a vector of activation weights and put through a sigmoid function f to create hidden layer vector $\mathbf{s}_j(t)$. Then, that hidden layer vector is multiplied by another vector of weights and put through a softmax function g in order to obtain outputs that define a probability distribution over the words in the vocabulary, which is the result for the timestep. This process continues for all timesteps (the length of the sequence of words). Taken directly from [25], Equations 5.1-5.3 describe a mathematical representations of each layer, input, hidden, and output, at a timestep:

$$\mathbf{x}(t) = [\mathbf{w}(t)^T \mathbf{s}(t-1)^T]^T \tag{5.1}$$

$$s_j(t) = f(\Sigma_i x_i(t) u_{ji}) \tag{5.2}$$

$$y_k(t) = g(\Sigma_j s_j(t) v_{kj}) \tag{5.3}$$

Backpropagation occurs through time in this model. This means that the error is propagated through a certain number of timesteps in order to update the weights of the network during training.

Lately, LSTM cells [26] have been used in recurrent neural networks like the one just described, and they have a stellar performance and language modelling while remaining efficient. Standing for long short term memory, they attempt to hold onto information that was seen an arbitrarily long number of cells in the past. Since humans producing natural language know about things that happened an arbitrary amount of time into the past, this knowledge influences what they will say next (e.g. which gendered pronoun should I use?) Briefly, LSTM cells accomplish this by keeping a cell state in each cell of a timestep. Gates choose which elements of this cell state to forget and remember based on previous hidden states and the current word, and a cell's hidden layer is based directly on this cell state. Practically, this means that LSTM cells are great at learning which information is important to keep track of based on context and the word itself.

Finally, I will describe input embeddings. In the RNNLM described above, recall that input layer represented words via a 1 of N encoding; this means that all words are equally different from each other. It is obvious that this is not the case; certain words are much closer in similarity than others. Input embeddings aim to describe the semantic meaning of each word as an n-dimensional feature vector. To create input embeddings, an easy way is to run a language model and map each word to a vector initialized randomly; allow the weights in these vectors to be adjusted during training by gradients. The resulting weights of all of these vectors for all of these words should now be decent approximations of their semantic meaning relative to each other. More complex methods exist for creating input embeddings [27], but that is the general idea. Pre-trained input embeddings exist which are trained on enormous amounts of data, and the more data one uses the better the input embeddings become. These pre-trained input embeddings are frequently used in tasks not even related to language modelling and have become almost necessary for models to be competitive with the state-of-the-art. All of the models in chapter 6 will use pre-trained FastText embeddings [28]. FastText embeddings have a special property that prevents words from being out-of-vocabulary because the embeddings are created from all of the subwords in a word; this technique is also assumed to make the embeddings more representative of a word's semantics and morphology.

CHAPTER 6: METHODS TO IMPROVE LOW-RESOURCE LANGUAGE PERFORMANCE ON NEURAL MODELS

This chapter will introduce methods that were used to try to improve the language modelling of low-resource creole languages. Each method utilizes concepts from linguistic theory and creole genesis about creole languages.

6.1 CROSS-LINGUAL PARAMETER SHARING

This method is derived from [7], which hypothesizes that since the hidden layers of a model are feature extractors and the last layer performs logistic regression (which is language-dependent), hidden layers are transferable between models of different languages. The experiments in the paper support this hypothesis, and it seems that certain languages work better than others when their parameters are shared. This is an intriguing result, but the reason as to why certain languages do better is not explored (which is what we will explore later for some models). We personally found that this method did not work well for language modelling of text, with some models consistently performing worse than the baseline that did not use cross-lingual hidden parameter layer sharing, so its results are not included. The method in the paper likely worked well because it used speech data and its input was represented as phonemes. Phonemes are absolutely more language independent (albeit not entirely - consider how many languages use clicks) than FastText input embeddings are. We assume that this method does not work well when the semantic spaces of the input embeddings are not perfectly aligned (which they will never be between two text languages unless they are trained together explicitly).

6.2 CROSS-LINGUAL NEURAL STACKING

This method is derived from [6] and [9]. Notably, the latter is described briefly in Chapter 3 and uses English as the language to support the target low-resource creole language of Singaporean English, Singlish (English is Singlish's lexifier). In contrast to the parameter sharing model, this model successfully allows us to lower perplexity on creole languages, so in chapter 8 we will study which high-resource languages allow this model to perform the best and why. As such, I will explain this model in detail here.

See figure 6.1. The first few layers of this model are from a pretrained language model of a high-resource language, but the output layer which performs a softmax is not included. This pretrained language model contains syntactic features extracted from this language, as

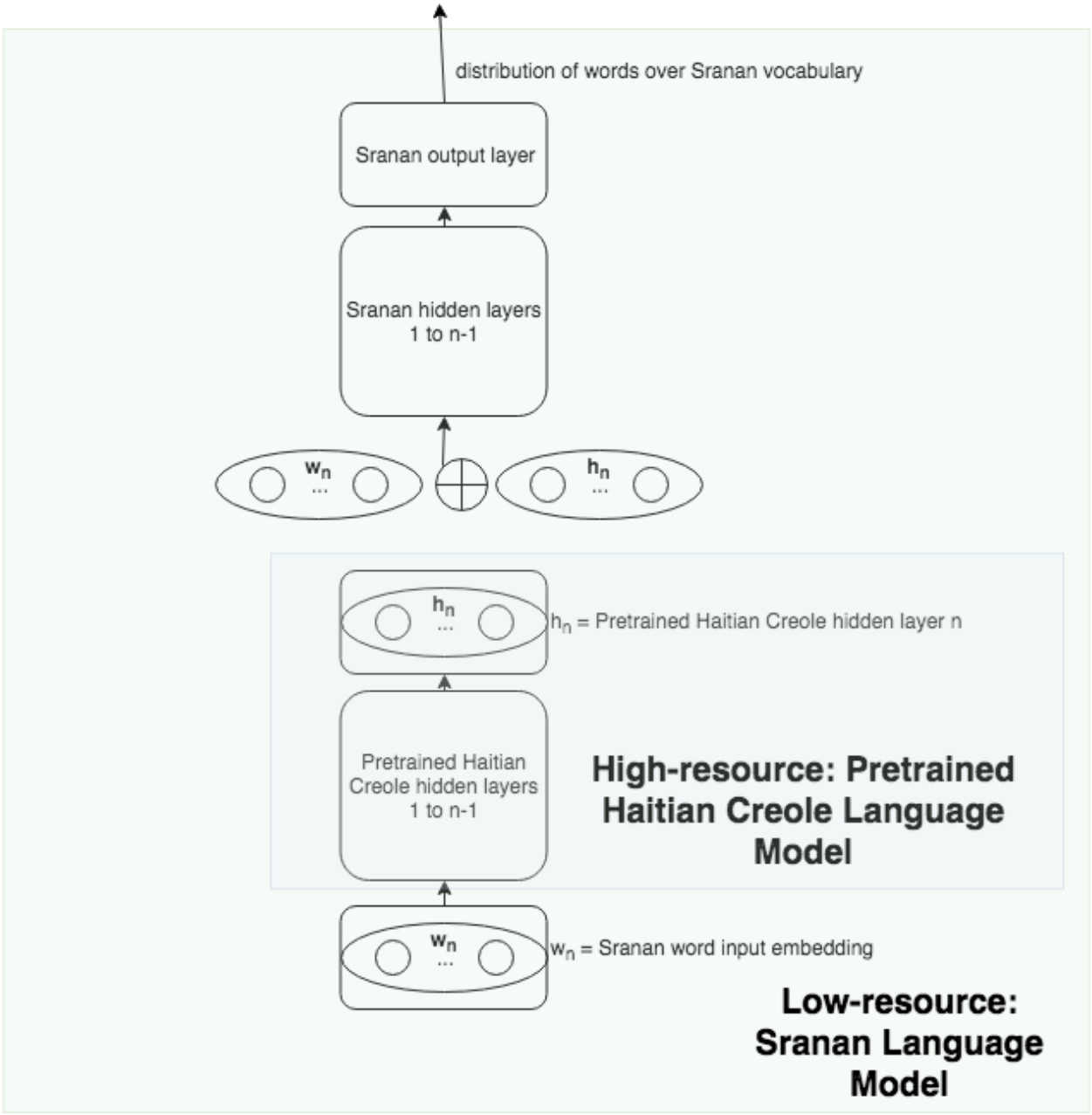


Figure 6.1: A Single Timestep of a Neural Stacking Language Model

any good language model should. See the following equations 6.1 and 6.2, recreated directly from [6]:

$$\vec{w}_a^i = tanh(\mathbf{W}^s \vec{h}^i) \tag{6.1}$$

$$\vec{r}_w^i = \vec{w}_c^i \oplus \vec{w}_a^i \tag{6.2}$$

Tanh of the final hidden vector from the high-resource pretrained model, $\vec{h}^i$, times a weight matrix $\mathbf{W}^s$ is known as is $\vec{w}_a^i$. Concatenate $\vec{w}_a^i$ with the original input embedding of the word $\vec{w}_c^i$, which gives us $\vec{r}_w^i$, a vector that effectively encodes all of the information from the pretrained high-resource language but also the input word. Afterward, a low-resource language is trained as normal but with this altered input embedding. Backpropagation occurs through the whole stacked model. It is called stacked because it effectively utilizes pretraining to combine two models into one, even if the models are trained on different languages.

In review, the steps to training this model are:

- Pretrain a language model for high-resource language (e.g. French) in its entirety
- Set the first layers of the neural stacking model to be the entirety of the pretrained high-resource language model, but do not include the output layer (which performs an operation to convert the output layer into a distribution over vocabulary words)
- The vector produced as the output at the last hidden layer, $\vec{h}^i$, will be multiplied by a matrix of weights, $\mathbf{W}^s$, and the tanh of it all will be taken to get $\vec{w}_a^i$ (see equation 6.1)
- The above vector is concatenated with the original input embedding for the word, $\vec{w}_c^i$, which will be used as the new input embedding, $\vec{r}_w^i$. This step effectively deeply encodes syntactic information from the high-resource language model into the low-resource language model
- The rest of the layers are a language model for the low-resource language, including a regular output layer
- Backpropagation occurs throughout the whole model, including the layers from the pretrained high-resource language model. The pretrained layers are allowed to change according to computed gradients

To review, [9] justified its choice of English to serve as the pretrained model in their stacking model due to its lexical and syntactic similarities to Singaporean English, the low-resource language of the model they wished to improve. Two things are striking about that claim:

- The lexical similarity of the two languages is not useful in the way that they appear to describe it is. Recall input embeddings: they turn word tokens into feature vectors representing the semantics of the word. The fact that Singaporean English establishes the majority of its lexicon based on English words doesn't really affect how well the neural stacking or any neural model will perform; that's why it's possible to use a language like Chinese at all in these cross-lingual model. Even though Chinese is lexically extremely different (logograms vs. phonograms), each Chinese logogram is still translated into the same input embedding vector as every other language. The only way in which a similar lexicon would matter for picking which model to use as the pretrained model in neural stacking is to say that if two languages have similar lexicons, the semantic space defined by the two languages should also be similar and can possibly be mapped but even this reasoning seems unsure.
- As we demonstrated in chapter 4, there does exist a syntactic similarity that is above average between creole languages and their lexifiers. The degree of syntactic similarity between creole and lexifier differs per creole. However, it is generally true that creole languages are much closer to each other than to any other languages, including their lexifier, as demonstrated in chapter 4 and [18]. For this reason, the syntactic similarity of English and Singaporean English is not so compelling when you see that languages like Haitian Creole, which has an exceptional amount of data available for a creole language, have much greater syntactic similarity.

As such, the experiments on neural stacking models in chapter 8 will focus on which pretrained languages lower the perplexity of creole language models the most and why. Specifically, similarity of lexicon and syntax will be examined as a possible factor in reducing perplexity.

6.3 PRE-TRAINED EMBEDDING ALIGNMENT

Recall the definition of a pre-trained embedding from chapter 5. Pre-trained input embeddings are an excellent way to enhance a neural model. However, each set of pre-trained embeddings describe different semantic spaces. For example, imagine that two languages X and Y have pre-trained embeddings trained on data in their respective languages. This means that if language X borrows a word from language Y, language X cannot simply borrow the pre-trained embedding for the word in language Y to describe the semantics of the word. This is because even though the words are the same, the embeddings for both languages are described in separate spaces because the embeddings were trained separately.

Consider the above two sections. Hidden parameter sharing directly transfers a high resource model's hidden layer parameters to initialize a low-resource language model. Since each language's input embeddings have a different semantic space, despite the hypothesis Ghoshal made for speech neural networks, this means that the hidden layer parameters also have to be in a different space this means that the features from the hidden layers cannot be directly transferred to a different language model. The same logic goes for the neural stacking model in Figure 6.1, a Sranan word embedding interacts with hidden layers that are trained to interact with Haitian Creole word embeddings. Unfortunately, Sranan and Haitian Creole have embeddings in different semantic spaces, so it will be difficult for the model to reconstitute the features learned by the Haitian Creole model during pretraining.

Consider what would happen if the embeddings were in the same semantic space. If this were true, it would follow that the hidden layers would describe the same features across models. This seems ideal we would be able to transfer knowledge from pretraining a separate language model much more easily with neural stacking! Thankfully, there is a method to do this defined by [29]. Given a list of vectors thought to have identical meanings in both languages, this method finds the transformation between the two vector spaces that will 'rotate' them into the same vector space. They find a rotation that minimizes the mean distance between a set of paired points (so that a pair of points will be words we have identified as having identical meanings in two languages). This transformation does not form a perfect mapping, but it definitely will allow a neural stacking model to be able to transfer the syntactic features it found during pretraining the high-resource language.

In chapter 8, we will align the high resource language's input embeddings to the low resource language's input embeddings. Since we want to test how neural stacking will decrease the perplexity of the model when different languages are used as a pretrained high resource model, aligning the embeddings of both models to the same semantic space will then mean that the two models are sharing syntactic features explicitly. It follows that if the syntactic features between the two languages are similar, if the high-resource and low-resource languages in a neural stacking model are more syntactically similar, the perplexity will be lower because more features were learned during pretraining.

6.4 SPELLING NORMALIZATION TO LEXIFIER

We will shift gears to discussing how to take advantage of lexifiers when attempting to improve creole language models. Lexifiers are known to provide the basis for the lexicon of their creole languages.

This idea is inspired directly from [2]. Creoles frequently do not have a standardized

orthography. As such, words that have the same semantic and contextual meaning have multiple spellings in datasets. Generally, the creoles' lexifiers tend to have a more standardized orthography, and the lexicons of the creoles and the lexifier highly match by definition. As such, this low-resource method is to normalize each word in the dataset so that it maps to the closest French word by minimizing edit distance. This will effectively limit the noise present in the variant spellings in the dataset. More sophisticated normalization techniques could result in less errors and lower perplexity.

6.5 USING THE LEXIFIER'S EMBEDDINGS DIRECTLY

A simple idea to experiment with is seeing if using the lexifier's embeddings instead of the creole's embeddings will yield better performance of the creole model. Since we are using FastText, the embeddings produced depend on n-grams. If the n-grams between lexifier and creole are similar enough and maintain enough semantic meaning between each other if they are similar, this could be very useful. This is because most lexifiers are extremely high resource languages due to European colonization (so languages like Portuguese, French, English, and Dutch are very common lexifiers), so their embeddings will be much more highly trained than a low resource creole's embedding. If the n-grams between the lexifier and the creole are not similar enough, the spelling normalization defined in the last section can be performed in tandem with this method to make the n-grams more similar at a potential loss to correct semantic meaning of n-grams.

CHAPTER 7: EXPERIMENTS

All experiments implement a word language model in PyTorch. Each model uses an LSTM with 2 hidden layers of dimension 200 and pre-trained 300-dimensional FastText embeddings at the input layer. Each model was trained for 20 epochs. The hyperparameters used are: an evaluation and test batch size of 10, a training batch size of 20, and a bptt length of 35. The learning rate was divided by 4 each time the perplexity does not decrease on the validiation set after an epoch. The test score is determined by the model saved at the epoch that achieved the lowest perplexity, and only the test scores are shown. The models were run on an Intel Xeon CPU E5-2640 v42.4Ghz processor, 512GB of memory, and GTX 1080 TI 11GB GPUs. Since there is very little data for the low resource languages in these experiments, the changes in perplexity are thus rather low. (None of these methods explicitly add data to the low resource language.) This is fine, though, because the low changes in perplexity are inevitable. It follows that the methods and suggestions in this section would accordingly cause more drastic drops in perplexity if there were more data.

7.1 DATASETS

Table 7.1 describes the languages in use. Only considered (except for Sranan and Lingala) if the language had a bible available at `https://github.com/christos-c/bible-corpus`, these languages were selected by examining some APiCS [16] morphosyntactic features that also exist on WALS [15], particularly the features shown in Table 4.1, and only using languages which have 30 or more of these features available.

In one experiment, our goal was to determine which language decreased perplexity of a low-resource creole language the most and to discover why. In order to validly compare the effects of language choice in neural stacking, the pretrained high-resource models

Afrikaans	Chinese	Farsi
Finnish	German	Greek
Hebrew	Hindi	Hungarian
Japanese	Kannada	Malagasy
Maori	Burmese	Russian
Spanish	Tagalog	Turkish
Vietnamese	English	French
Haitian Creole	Sranan	Lingala

Table 7.1: Languages used

must have equivalent amounts of data. As such, we used translations of the bible in each language in order to get about equal amounts of data per pretrained model. Unfortunately, in the case of Sranan and Lingala, only New Testament translations could be found at `http://ebible.org/pdf/srnNT/` and `http://gospelgo.com/f/lingala_nt.htm`. All other bible translations are found here: `https://github.com/christos-c/bible-corpus`. These bibles and new testaments were then also used for other experiments. The data is preprocessed and tokenized with the same script used to create the pretrained FastText embeddings. The data is split 80/10/10 into a training, validiation, and test set respectively. The data was split per book so that 80% of each book in the bible is in the training set and 10% of each book in the bible is in the validation set.

7.2 NEURAL STACKING

The experiments in this section will attempt to identify the factor that explains which high-resource pretrained language model will aid a low-resouce language in lowering its perplexity the most via neural stacking. In this section, all neural stacking models' pretrained language model use pretrained FastText embeddings that are aligned to the low resource model; as explained in chapter 6, this will help the neural stacking model directly transfer the syntactic features it learned during training to the low resource model. The factors examined are syntactic similarity and lexical similarity. In this section, all tokens that cannot be resolved to a pretrained embedding become the same token with the same embedding, '<UNK>', which is a special token that represents rare words.

7.2.1 Syntactic and Lexical Similarity

Define $\mathbf{v}$ and $\mathbf{w}$ as vectors that represent values each language has for all word order features from the APiCS-WALS database. The features that define word order in the APiCS-WALS database are features 1, 2, 3, 4, 5, 6, 7, 8, and 12 as defined in Table 4.1. Throw out all entries in $\mathbf{v}$ and $\mathbf{w}$ for which one vector has missing a value. We define syntactic similarity between two languages as:

$$ssim(\mathbf{v}, \mathbf{w}) = \frac{|\mathbf{v}| - H(\mathbf{v}, \mathbf{w})}{|\mathbf{v}|} \tag{7.1}$$

Word order is very relevant to language modelling, which explicitly must learn word order and thus must encode them as features.

Define $\mathbf{y}$ and $\mathbf{z}$ as vectors that represent which cognate type each language has for all

200 words on the Swadesh list (read [22] for more clarification on this terminology). If two languages share a cognate, their values for the Swadesh list word in this vector are identical; otherwise the values are not identical. We define lexical similarity between two languages as:

$$lsim(\mathbf{y}, \mathbf{z}) = \frac{|\mathbf{y}| - H(\mathbf{y}, \mathbf{z})}{|\mathbf{y}|} \tag{7.2}$$

This is also known as the lexicostatistical percentage in [22], which is also the paper from which I collect these similarities.

7.2.2 Do syntax and lexical similarity increase as perplexity decreases?

Figures 7.1, 7.2, and 7.3 show the graphs of syntactical similarity of a language pair vs. the difference of the baseline on the low-resource language alone minus the neural stacking model on the language pair. In other words, each point on this graph denotes how much a language pair reduced the perplexity in its neural stacking model where the creole language being plotted is the low-resource language, and the other language being plotted is the high-resource language that is pretrained.

Figures 7.4 and 7.5 show the same, but the x-axis is instead lexical similarity based on lexicostatistical percentages from [22]. If the high resource language was not Indo-European, the lexical similarity is set to 0. Since Lingala does not have a Indo-European lexifier, there was no point in showing its graph because its lexical similarity would always be 0.

In all of the graphs of syntactic similarity, 7.1-7.3, there is a positive correlation between difference in perplexity and word order / syntactic similarity. This means that the perplexity is reduced further when using a neural stacking model if the pretrained model has a more similar word order to the creole language! It appears to always be a good idea to choose a language that is more syntactically similar when creating a neural stacking model.

In the graphs of lexical similarity, 7.4 has a negative correlation and 7.5 has positive correlation. For figure 7.4, this means that for Haitian Creole, choosing a syntatically similar language is vastly superior to choosing a lexically similar language. For figure 7.5, choosing a lexically similar language is only slightly less superior to choosing a syntactically similar language; which is reasonable, because lexically similar languages and lexifiers are somewhat syntactically close as well as demonstrated in chapter 4. This shows that choices that are lexically closer to a creole language are not necessarily bad, but as shown in figure 7.4, for some creoles lexically close languages can be awful at reducing perplexity; it just depends on the individual language.

Tables 7.2-7.4 reiterate the same information in a different format. The same values

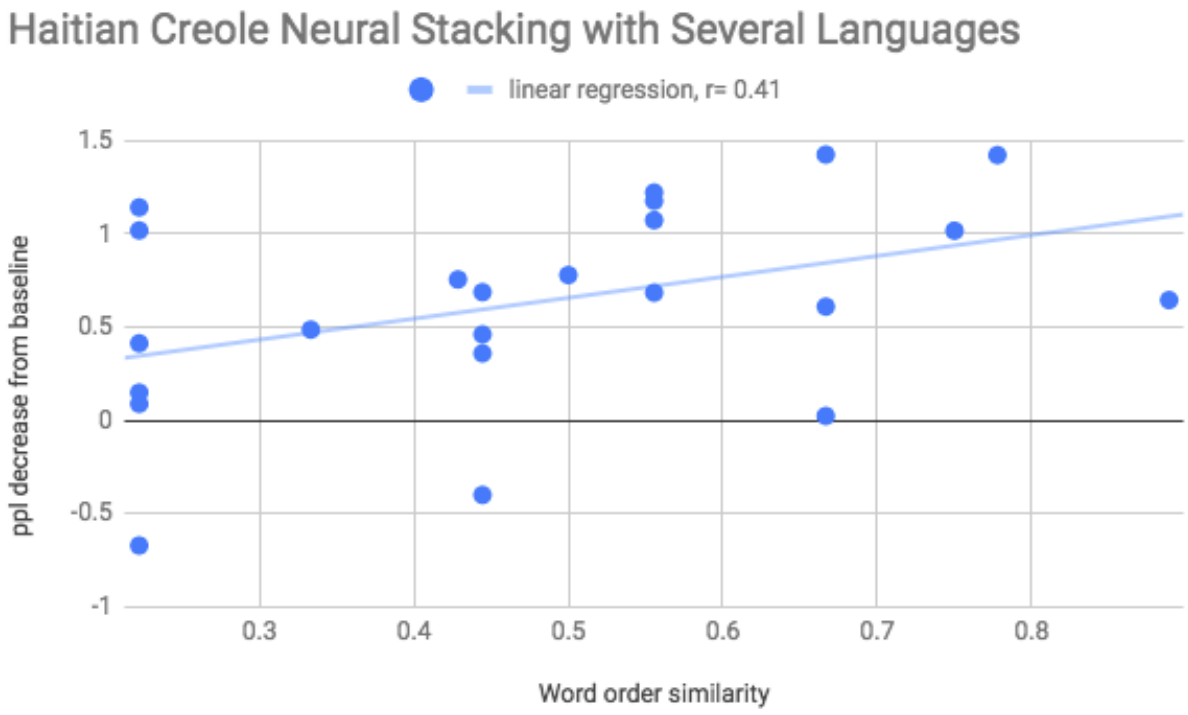


Figure 7.1: Haitian creole's perplexity reduction vs. syntactic similarity

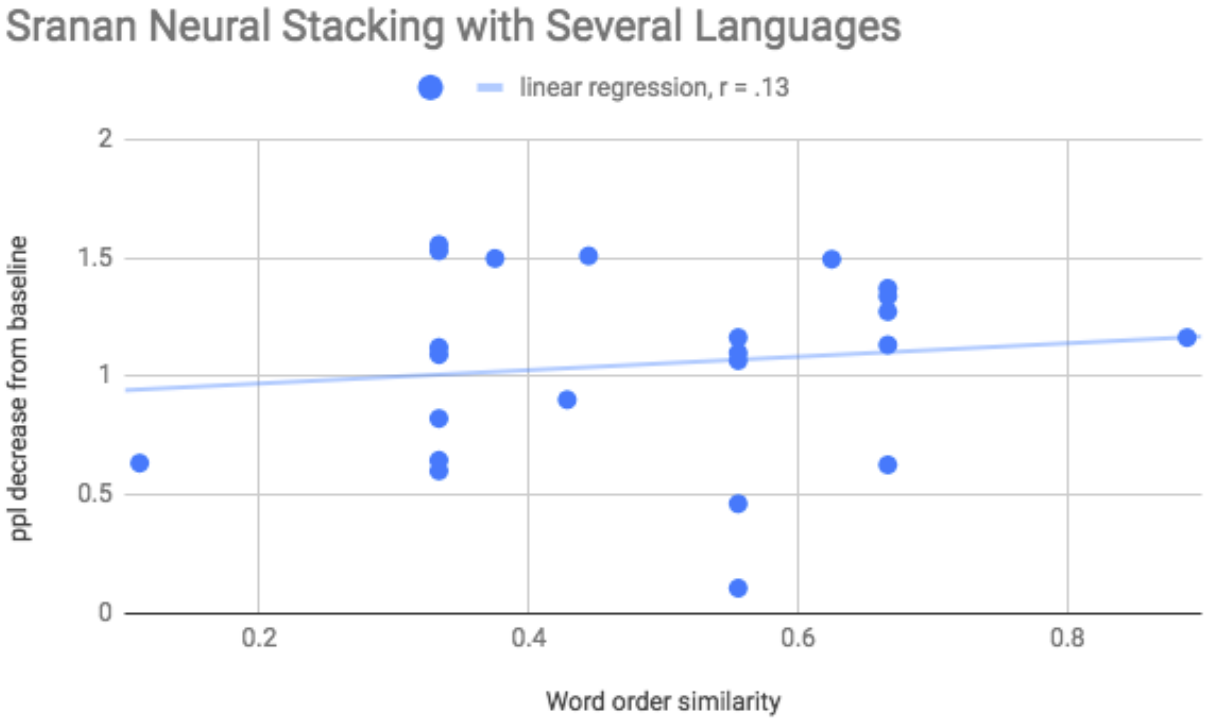
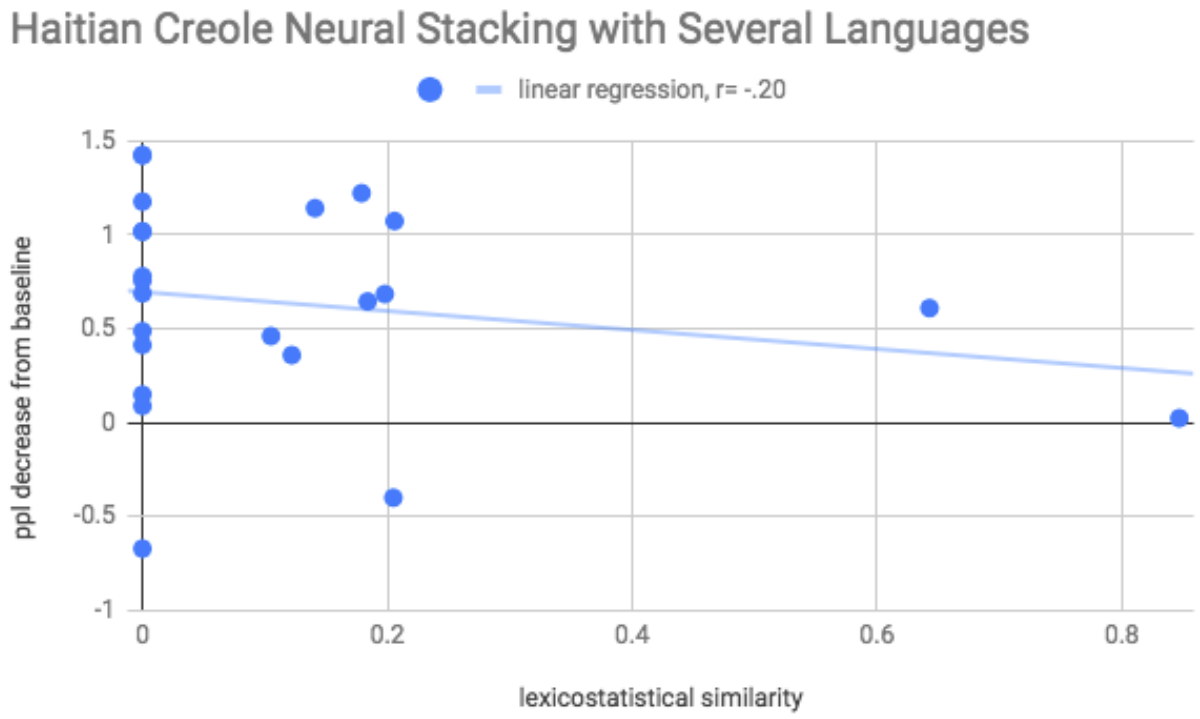
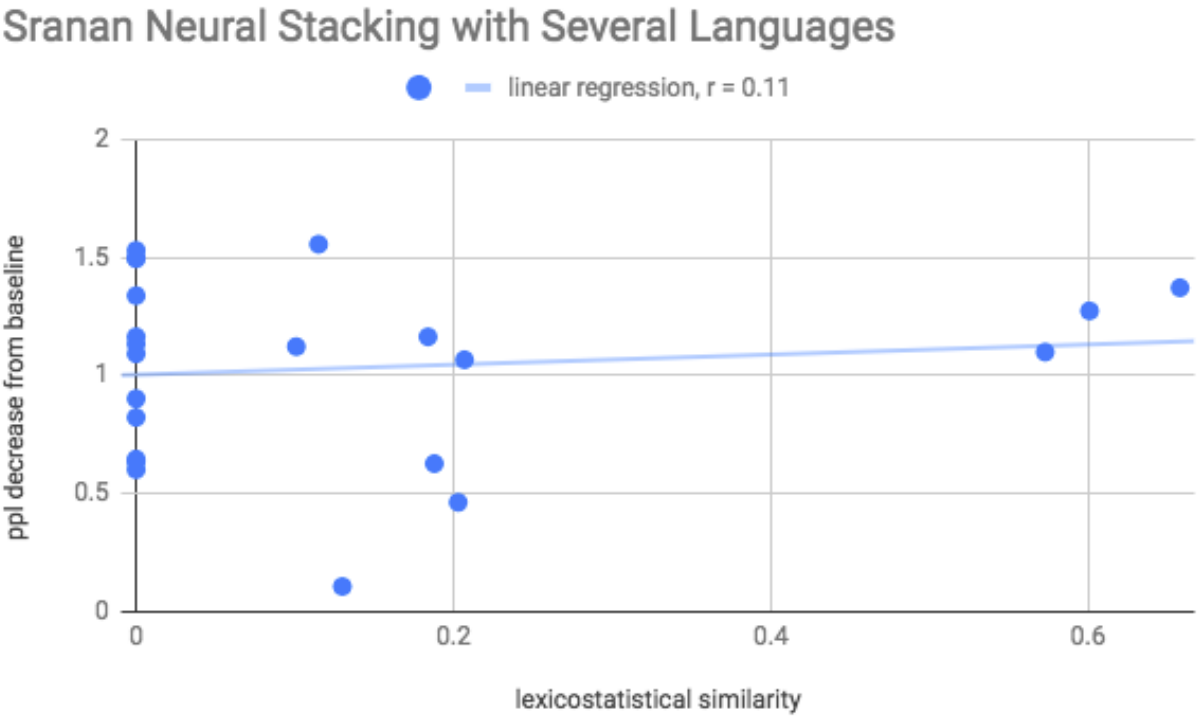


Figure 7.2: Sranan's perplexity reduction vs. syntactic similarity



.png

Figure 7.3: Haitian creole's perplexity reduction vs. lexical similarity



.png

Figure 7.4: Sranan's perplexity reduction vs. lexical similarity

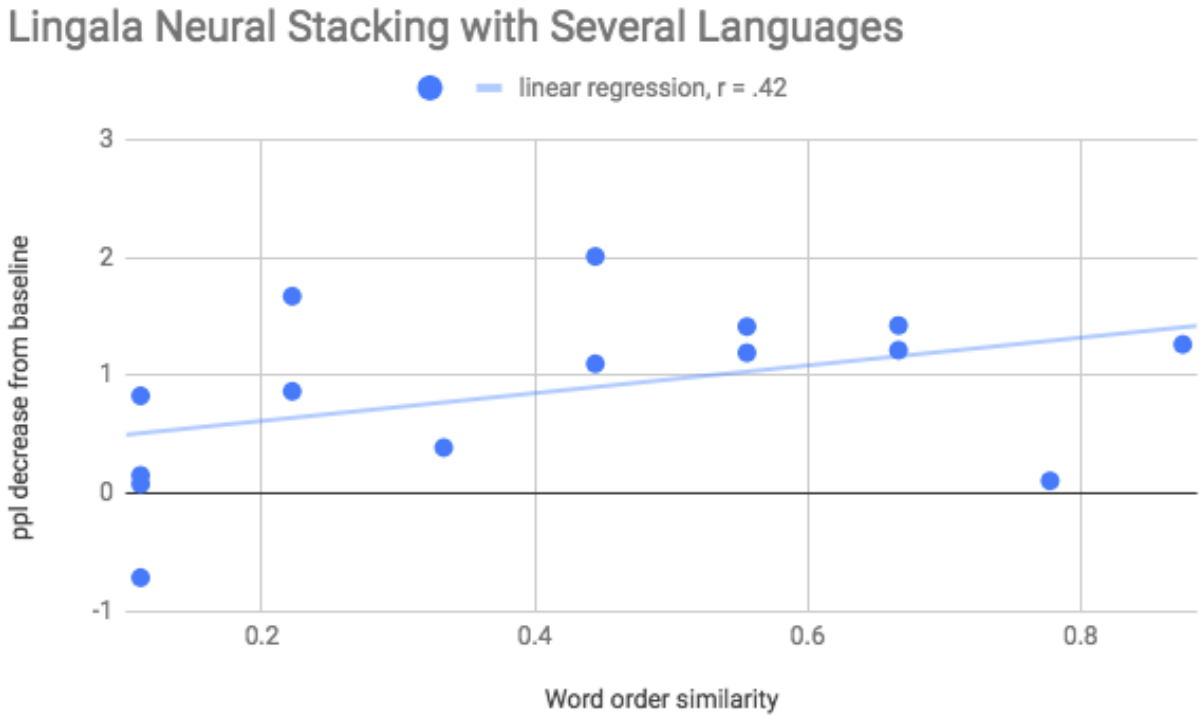


Figure 7.5: Lingala's perplexity reduction vs. syntactic similarity

	Avg. ppl reduction	Syn. similarity	Lex. similarity
Creoles (w Afrikaans)	**1.05**	0.70	0.13
Creoles (w/o Afrikaans)	1.04	0.78	0.09
Romance langs	0.32	0.67	0.75
Indo-European langs	0.58	0.54	0.28
All langs	0.63	0.48	0.12

Table 7.2: Haitian Creole Neural Stacking by Group

	Avg. ppl reduction	Syn. similarity	Lex. similarity
Creoles (w Afrikaans)	1.20	0.70	0.26
Creoles (w/o Afrikaans)	1.16	0.72	0.09
Germanic langs (w Afrikaans)	**1.25**	0.63	0.61
Germanic langs (w/o Afrikaans)	1.23	0.611	0.62
Indo-European langs	0.98	0.58	0.30
All langs	1.05	0.49	0.13

Table 7.3: Sranan Neural Stacking by Group

	Average ppl reduction	Word order similarity
Creoles (w Afrikaans)	**1.62**	0.56
Creoles (w/o Afrikaans)	1.42	0.61
All langs	0.87	0.41

Table 7.4: Lingala Neural Stacking by Group

for perplexity are examined, but this time they are grouped and averaged alongside the group's average similarity metrics. Note that Afrikaans is singled out because it is both Germanic and partially considered a creole language.

In table 7.2, the grouping with all creoles yields the highest average perplexity reduction. This group also has a very low lexical similarity and a very high word order/syntactic similarity, which further supports the fact that similar syntax reduces perplexity but not having lexical similarity does not hinder it. In fact, languages in the same language family for Haitian Creole actually did not reduce perplexity as much as languages in other families. In table 7.3, Germanic languages did the best at reducing perplexity slimly followed by creole languages. This shows that sometimes lexifiers and languages in the lexifiers' family can improve the model's performance as well. Note that creoles did very well despite only having a rather low 0.26 lexical similarity on average; it's likely due to the high syntactic similarity. In table 7.4, creoles lessen reduction more than all languages on average do while maintaining a higher syntactic similarity again.

7.2.3 Why examine this for creole languages?

There is a subtle reason for why examining this data for creole languages as the low-resource language in neural stacking is more interesting than for any other language. It is true that these results will likely generalize to non-creoles, but keep this in mind: for languages that undergo single parent evolution, if two pairs of languages have similar lexicon, it is very likely that the two pairs of languages also have a similar syntax. As examined in chapter 3 and 4, we know that this is not the case for creole languages. Due to their odd genesis, there are languages for which they are lexically close to but not as syntactically close to (lexifiers and languages in their family). Similarly, there are languages for which they are extremely syntactically close to but are not lexically close to at all (other creole languages). As such, the choice of which language to use for creole languages is more complex given these two factors.

Technique	Reduction in ppl from baseline
Use French embeddings.	0.69
Normalize to French, use Haitian Creole embeddings.	**0.84**
Normalize to French, use French embeddings.	-0.62

Table 7.5: Using French to decrease the perplexity of a Haitian Creole model

Technique	Reduction in ppl from baseline
Use English embeddings.	0.54
Normalize to English, use Sranan embeddings.	0.23
Normalize to English, use English embeddings.	**0.79**

Table 7.6: Using English to decrease the perplexity of a Sranan model

7.2.4 Analysis

Other creole languages as a group are the best or very close to the best at reducing perplexity for every low resource creole. Keep in mind that this is while two of the three creoles, Sranan and Lingala, have less than half the data of all other languages because only a New Testament is available for them. Creoles tend to have very high syntactic similar but very low lexical similarity, which leads us to believe that syntax is the most important factor in multilingual neural models. As we mentioned in the above section, creole languages gave us the unique opportunity to decouple lexical similarity and syntactic similarity to examine which one is truly the most salient. Finally, we propose that high resource languages in a multilingual model should be syntactically close in order to best increase performance. This is not a hard rule, as there are still data points on Figures 7.1-7.5 that have low syntax similarity but reduced perplexity greatly. Nonetheless, the correlations with syntax and perplexity reduction are consistent across languages while the correlations with lexical similarity and perplexity reduction are not.

7.3 SPELLING NORMALIZATION AND ALTERNATE EMBEDDINGS

Instead of using <UNK> tokens like the previous section, the models in these sections assign out of vocabulary words random embeddings. Despite lexical similarity not being so important in terms of choice of language in a multilingual neural model, lexifiers can still be used in interesting ways like described in chapter 6. Table 7.5 and 7.6 hold the results of spelling normalization to a lexifier with and without using lexifier embeddings. It also holds the results of just using lexifier embeddings with spelling normalization.

In table 7.5, the best reduction comes from normalizing words to French from Haitian Creole and then using Haitian Creole embeddings. In table 7.6, the best reduction comes from normalizing words to English from Sranan and then using English embeddings. Normalization seems to help in both cases by reducing variance in spelling of items with similar semantics. However, why does Haitian Creole prefer using its own embeddings over a lexifier? The answer is simple: FastText embeddings are trained on Wikipedia articles, and there are *a lot* of articles for Haitian Creole. In fact, there are fifty times more articles written in Haitian Creole than there are in Sranan. Haitian Creole's embeddings appear to be good enough such that using another language's embeddings, no matter how good and similar they are, does not help (in fact, it does worse than the baseline the Haitian Creole embeddings help that much). However, in Sranan's case, only 1000 articles are written in Sranan in Wikipedia, so its embeddings are pretty bad. That is why normalizing the language and using English embeddings helps not only does it reduce noise in the Sranan vocabulary, but it gets to use the much superior English embeddings well. All in all, utilizing lexifiers is a useful multilingual approach for creating better language models.

CHAPTER 8: CONCLUSION

Multilingual neural models give languages with little data a chance to obtain better performance by taking advantage of the features that a model in another language has learned by pretraining. In order to transfer these features effectively, we have found that using a neural stacking method with embeddings in the same semantic space works well for the task of creating a language model. Furthermore, through examining low-resource creole languages' typology and how they are similar and dissimilar to other languages, we came to the conclusion that syntactic similarity is the best metric for choosing a high resource model to pretrain for these multilingual models, and not so much lexical similarity. Of course, other factors have to be taken into account besides syntactic similarity, like the amount of data available to train the high resource language's pretrained model, but the fact that creole languages Sranan and Lingala acting as the 'high resource language' with half the amount of data of other more dissimilar languages could perform as well or better at reducing perplexity shows that syntactic similarity is a strong factor in how well these multilingual neural language models will perform.

Lexical similarity should not be totally disregarded, however. Once again, by examining how low-resource creole languages performed in terms of relative perplexity difference, normalizing the spelling of languages with less standardized orthography helped decrease language model perplexity when creole languages were normalized to vocabulary in their lexifier. In the case of Sranan, which had poorly trained embeddings due to lack of resources, spelling normalization in combination with using its lexifier's embeddings increased performance even more.

Creole languages are fascinating and deserve to be more well-studied computationally. Much is unknown or unsure about them still, but uncovering their mysteries can tell us a lot about how language evolution and the universal language principles of the human mind operate.

REFERENCES

[1] Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

[2] S. Hewavitharana, N. Bach, Q. Gao, V. Ambati, and S. Vogel, "CMU Haitian Creole-English translation system for WMT 2011," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 386–392.

[3] R. Hwa, P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak, "Bootstrapping parsers via syntactic projection across parallel texts," *Natural language engineering*, vol. 11, no. 3, pp. 311–325, 2005.

[4] O. Zennaki, N. Semmar, and L. Besacier, "Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks," in *29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 2015.

[5] C. Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *CoRR, abs/1503.03535*, vol. 15, 2015.

[6] H. Chen, Y. Zhang, and Q. Liu, "Neural network for heterogeneous annotations," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016. [Online]. Available: https://aclweb.org/anthology/D16-1070 pp. 731–741.

[7] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[8] R. Dabre, A. Sukhoo, and P. Bhattacharyya, "Anou Tradir: Experiences In Building Statistical Machine Translation Systems For Mauritian Languages–Creole, English, French," in *Proceedings of the 11th International Conference on Natural Language Processing*, 2014, pp. 82–88.

[9] H. Wang, Y. Zhang, G. L. Chan, J. Yang, and H. L. Chieu, "Universal Dependencies Parsing for Colloquial Singaporean English," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, 2017, pp. 1732–1744.

[10] C. Lefebvre, "Relexification in creole genesis revisited: The case of Haitian Creole," *Substrata versus universals in creole genesis*, pp. 279–300, 1986.

[11] L. Campbell, *Historical linguistics*. Edinburgh University Press, 2013.

[12] M. Baptista, "New directions in pidgin and creole studies," *Annu. Rev. Anthropol.*, vol. 34, pp. 33–42, 2005.

[13] S. Mufwene, "Creolization is a social, not structural, process," *Degrees of restructuring in Creole languages*, pp. 65–84, 2000.

[14] J. H. McWhorter, "Identifying the creole prototype: Vindicating a typological class," *Language*, pp. 788–818, 1998.

[15] M. S. Dryer and M. Haspelmath, Eds., *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. [Online]. Available: https://wals.info/

[16] S. M. Michaelis, P. Maurer, M. Haspelmath, and M. Huber, Eds., *APiCS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. [Online]. Available: https://apics-online.info/

[17] Y. Murawaki, "Statistical modeling of creole genesis," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 1329–1339.

[18] A. Daval-Markussen and P. Bakker, "Explorations in creole research with phylogenetic tools," in *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Association for Computational Linguistics, 2012, pp. 89–97.

[19] M. Parkvall, "The simplicity of creoles in a cross-linguistic perspective," *Language complexity: Typology, contact, change*, pp. 265–285, 2008.

[20] S. S. Mufwene, "The universalist and substrate hypotheses complement one another," *Substrata versus universals in creole genesis*, pp. 129–162, 1986.

[21] A. Stamatakis, "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models," *Bioinformatics*, vol. 22, no. 21, pp. 2688–2690, 2006. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btl446

[22] I. Dyen, J. B. Kruskal, and P. Black, "An Indoeuropean classification: A lexicostatistical experiment," *Transactions of the American Philosophical society*, vol. 82, no. 5, pp. iii–132, 1992.

[23] L. Nakhleh, T. Warnow, D. Ringe, and S. N. Evans, "A comparison of phylogenetic reconstruction methods on an Indo-European dataset," *Transactions of the Philological Society*, vol. 103, no. 2, pp. 171–192, 2005.

[24] H. Schwenk, D. Dchelotte, and J.-L. Gauvain, "Continuous space language models for statistical machine translation," in *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006, pp. 723–730.

[25] T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[27] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[28] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *arXiv preprint arXiv:1607.04606*, 2016.

[29] S. L. Smith, D. H. Turban, S. Hamblin, and N. Y. Hammerla, "Offline bilingual word vectors, orthogonal transformations and the inverted softmax," *arXiv preprint arXiv:1702.03859*, 2017.