Detecting Loanwords in Emakhuwa: An Extremely Low-Resource Bantu Language Exhibiting Significant Borrowing From Portuguese

Felermino D. M. A. Ali[1,2,3], Henrique Lopes Cardoso[1,2], Rui Sousa-Silva[3,4]

[1]Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC / LASI)

[2]Faculdade de Engenharia da Universidade do Porto (FEUP),
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

[3]Centro de Linguística da Universidade do Porto (CLUP)

[4]Faculdade de Letras da Universidade do Porto,
Via Panorâmica, 4150-564 Porto, Portugal

{up202100778, hlc}@fe.up.pt, rssilva@letras.up.pt

Abstract

The accurate identification of loanwords within a given text holds significant potential as a valuable tool for addressing data augmentation and mitigating data sparsity issues. Such identification can improve the performance of various natural language processing tasks, particularly in the context of low-resource languages that lack standardized spelling conventions. This research proposes a supervised method to identify loanwords in Emakhuwa, borrowed from Portuguese. Our methodology encompasses a two-fold approach. Firstly, we employ traditional machine learning algorithms incorporating handcrafted features, including language-specific and similarity-based features. We build upon prior studies to extract similarity features and propose utilizing two external resources: a Sequence-to-Sequence model and a dictionary. This innovative approach allows us to identify loanwords solely by analyzing the target word without prior knowledge about its donor counterpart. Furthermore, we fine-tune the pre-trained CANINE model for the downstream task of loanword detection, which culminates in the impressive achievement of the F1-score of 93%. To the best of our knowledge, this study is the first of its kind focusing on Emakhuwa, and the preliminary results are promising as they pave the way to further advancements.

Keywords: Loanword detection, low-resource settings, Emakhuwa, African languages

1. Introduction

Loanwords are words taken from one language and then incorporated into another language's vocabulary (Kang, 2011). In general, when multiple languages coexist in the same community, they commonly exchange and borrow words from one another, resulting in a mutual enrichment of their respective vocabularies. This phenomenon can be observed in various African languages, where borrowing words from colonial-adopted languages is common, enabling lexicon enrichment with culturally foreign terms.

However, challenges arise when adapting borrowed words into primarily spoken languages that lack standardized spelling. Existing corpora in African languages reveal significant spelling inconsistencies, some of which are associated with inconsistent lexical borrowing (Adebara and Abdul-Mageed, 2022). While these inconsistencies may go unnoticed in spoken conversations due to similar pronunciation, they contribute to poor-quality textual data when encountered in written form. This research presents a method for the automated detection of these borrowed words in Emakhuwa.

Emakhuwa, also known as Makua, Macua, or Makhuwa, is a Bantu language spoken by over 7 million people in northern and central Mozambique. It is the most widely used language in Mozambique, surpassing even the number of speakers of Mozambique's official language, Portuguese.

Existing Emakhuwa's text corpora exhibit a significant presence of loanwords from Portuguese, often written using inconsistent spelling (Ali et al., 2021). This happens due to the absence of certain sounds in Emakhuwa.

There are three main ways borrowed words are adapted from a donor into the recipient language (Kang, 2011):

- Phonetic adaptation: The word is made to sound as close as possible to the donor pronunciation.
- Phonotactic adaptation: The word is changed to follow the recipient's language sound patterns.
- Unchanged borrowing: The recipient word is kept identical to the donor.

To illustrate this, in Portuguese-Emakhuwa translation, the term *rádio* (radio in English), could

LREC-COLING 2024, pages 4750–4759

potentially be translated using eighteen different spellings, all of which are arguably valid: *rádio*, *rádiyo*, *rádiyu*, *rádio*, *rátiyo*, *rátiyu*, *radio*, *radiyo*, *radiyu*, *ratio*, *ratiyo*, *ratiyu*, *raadio*, *raadiyo*, *raadiyu*, *raatio*, *raatiyo*, *raatiyu*.

These variations primarily arise from the adaptation of specific Portuguese sounds, such as *rá*, *d*, and *o*, into Emakhuwa . However, it is important to note that this number of spellings could increase further due to Emakhuwa being an agglutinative language. Therefore, the word *radio* in Emakhuwa could be preceded or followed by different prefixes and suffixes, resulting in an even greater variety of spellings. For example, in *mu-ratio-ni*, which means "inside the radio" (i.e., "radio office"), the prefix *mu* and the suffix *ni* is added to the base word *ratio*.

Training Natural Language tools using such data, particularly machine translation systems, become challenging since spelling inconsistency potentially exacerbates the out-of-vocabulary word problem.

This study presents a novel approach to detecting loanwords in Emakhuwa borrowed from Portuguese. This is the first study of its kind specifically targeting the Emakhuwa and makes the following contributions:

- A dataset containing 8,055 loanwords identified in Emakhuwa borrowed from Portuguese.
- A supervised method for loanword identification, which achieved a performance of 93% F1-score. Notably, our approach can detect loanwords regardless of spelling variations. We achieve this by incorporating a sequence-to-sequence model during the feature extraction and fine-tuning the CANINE (Clark et al., 2021) model on the proposed dataset.

Our proposed method shows promising findings for potential application to language pairs in which one language benefits from enriching its vocabulary through borrowing from another language, particularly when the recipient language has agglutinative characteristics.

We make our loanword dataset and source code publicly available to foster further research at the EmakhuwaNLP repository[1].

2. Related Work

Lexical borrowing has received little attention in Natural Language Processing (NLP), despite its potential applications in various NLP tasks. One notable application is Machine Translation (Nath et al., 2022), where lexical borrowing can address the out-of-vocabulary (OOV) problem. Studies such as Mi et al. (2020, 2021) have also suggested using loanwords to improve machine translation and handle co-referents and named entities, potentially leading to enhanced translation quality (Ortega et al., 2021; Nath et al., 2022). Additionally, loanword detection has shown promising results in phylogenetic reconstruction, particularly in cognate detection (List and Forkel, 2022a).

However, identifying loanwords presents a significant challenge. A vast majority of previous studies has approached the loanword detection task as a classification problem, training models using similarity features such as phonetic, spelling, and semantic characteristics between the donor and recipient words (Mi et al., 2014, 2018, 2020, 2021; Miller et al., 2021; Nath et al., 2022; Miller and List, 2023). Phonetic and spelling similarities are relatively easy to compute using classical edit distance methods (Levenshtein, 1965). However, measuring semantic similarity poses challenges in low-resource languages due to the lack of pre-trained vector representations (e.g., word embeddings), which are essential for semantic analysis. To address this limitation, Mi et al. (2018), Mi (2023), and Nath et al. (2022); Mi (2023), proposed leveraging multilingual language models to extract cross-language proximity of contextualized word embeddings, enabling loanword detection through cosine similarity.

Looking at it from a different perspective, (Miller et al., 2021; List and Forkel, 2022b) tackle the loanword detection problem using a wordlist, which incorporates the modeling of phonology and phonotactics. Miller et al., 2021 focused on monolingual wordlist modeling with classifier-based methods, whereas Miller et al., 2021 focused on multilingual wordlist modeling with cognate-based methods. Building upon their previous works, Miller and List, 2023 further advance their research by comparing their cognate-based method against closest match and classifier-based methods. Interestingly, their findings indicate that the classifier-based approach outperforms the other methods in loanword detection.

Previous works on loanwords in low-resource settings often overlooked scenarios involving languages with spelling variations. This is particularly relevant for primarily spoken languages, where the graphemic transcription of loanwords may result in inconsistent spelling patterns in written text. Therefore, the primary objective of this study is to develop a robust model capable of identifying loanwords within a given text, even when they exhibit inconsistent spelling patterns. This challenge arises due to the lack of direct one-to-one correspondence between the donor and recipient word. Conventional loanword detection methods rely on supervised training, requiring the provision of recipient words, donor words, and corresponding labels. Our pro-

[1] https://github.com/felerminoali/emakhuwa-nlp/tree/master/datasets/loanwords

posal departs from this approach by omitting donor information in training and inference. This means the model trains only on recipient language words, making it more practical for real-world situations where identifying loanwords often happens without knowing the source language.

3. Data Collection

The creation of the dataset for our loanword detection model involved a manual effort, which included the construction of a loanword dataset and a bilingual dictionary, as detailed below.

3.1. Loanword Dataset

The dataset was collected manually by volunteers fluent in both Emakhuwa and Portuguese. They contributed to collecting data for training consisting of 8,055 examples of loanwords. Since there is no existing literature to support us on the proportion of loanwords from Portuguese into Emakhuwa, we opted to create a balanced distribution dataset. Thus, we constructed a dataset with an approximate distribution of positive examples (loan) and negative examples (not loan).

The negative examples were also manually collected to assemble a robust dataset for training a resilient model to nuances between Emakhuwa and Portuguese. To achieve this, we gathered the following types of negative examples:

- Hard negatives as defined by Nath et al., 2022: Non-loanwords words that are phonetically similar to a word in the donor language, making them challenging to identify as non-loans. For instance, *vaale* is an Emakhuwa native meaning "there" but sounds close to *vale*, which is a Portuguese word for "valley." We collected a total of 2,234 examples of hard negatives.

- Random: In this category, we randomly sampled words from Emakhuwa's word list, which native speaker later validated. This category includes of 5,325 examples.

After gathering the data, we partitioned it into the training and testing sets, as illustrated in Table 1.

Table 1: Dataset partitions

Partition	Class	Amount
Train	Positives	7555
	Hard negatives	2234
	Negatives	5325
Test	Positives	500
	Hard negatives	500
	Negatives	500

3.2. Loanword Dictionary

We also composed a bilingual lexicon dictionary of loanwords by aligning the donor and recipient words (see Table 2). The resulting dictionary and dataset are the foundational training data for our loanword detection model.

Table 2: Sample of the bilingual loanword dictionary

Donor (Portuguese)	Recipient (Emakhuwa)
especialista	*espesiyaalista*
online	*olaayini*
Tribunal	*etiripuunale*
Tribunal	*itiripuunale*
Tribunal	*itribunaale*
poções	*ipoosawu*
outubro	*utupuru*
açúcar	*esukhiri*
açúcar	*esukhari*
semana	*esumana*

4. Feature Extraction

Nath et al. (2022) inspires our feature extraction methodology. However, we have slightly adjusted the feature extraction method. In contrast to Nath et al. (2022), our method abstains from considering the donor counterpart when extracting these features. Instead, we use as an auxiliary resource an online dictionary containing all words from donor languages. We hypothesize that identifying loanwords without prior knowledge of the borrowed word would make the model flexible to spelling variations and more resilient to language change.

We use two categories of features to identify loanwords: *language-specific* and *similarity-based*. Language-specific features were extracted by leveraging our knowledge of the languages under consideration. Conversely, similarity-based features are determined by using lexicons and phonetic measurements.

4.1. Language-specific features

Language-specific features were derived from our in-depth understanding and knowledge of Emakhuwa and Portuguese. By carefully analyzing each language's specific linguistic traits and patterns, we are able to discern unique lexical markers indicating the presence of loanwords. These features are as follows:

- **Foreign Letters**: words in Emakhuwa should be written using Emakhuwa's alphabet. Thus, if the word contains any letters that do not belong to the Emakhuwa alphabet (see Table 3), we assign this feature a value of "1" (True); otherwise, the value is "0" (False). Some examples of letters from the Portuguese alphabet that are typically not used in Emakhuwa are "b", "d", "g", "j", "z", "q", "ç", "â", "ã", "ê", "ô", "õ".

- **Emakhuwa letters**: Emakhuwa, as a tonal language, uses specific letters to represent different tones. These include long vowels such as "aa" "ee" "ii" "oo" and "uu", the alphabet, as well as 23 other graphemes allowed in the language: "fy", "kh", "kw", "khw", "lw", "ly", "mw", "my", "ph", "pw", "py", "phy", "phw", "rw", "ry", "sy", "th", "thw", "tt", "tth", "ttw". If these graphemes appear in the word, we assign this feature a value of "1"; otherwise, the value assigned is "0".
- **Affixes**: Emakhuwa, as an agglutinative language, uses affixes in conjunction with a radical. For instance, prefixes like "ki-" and "kin-" mark the infinitive tense of specific verbs. However, these prefixes, among others, are exclusively applied to native Emakhuwa words and are not used in borrowed terms. Similarly, native Emakhuwa words incorporate verbal suffixes, with some of these suffixes being less common in loanwords, such as "-ela," "-aka," "-iwa," and so on. We have compiled a comprehensive list of the most prevalent prefixes and suffixes found in native Emakhuwa words. This process has resulted in a list of 96 distinct prefixes and suffixes. So, for this feature, We assign a value of 1 to indicate the presence of these affixes in a target word and 0 if they are absent.
- **Adjacent Consonants**: In Emakhuwa, adjacent consonants within a sequence are not grammatically allowed. To identify violations of this rule, we assign this feature a value of "1" if a sequence contains adjacent consonants; otherwise, the value is "0", meaning the absence of adjacent consonants. Emakhuwa's letters, however, are an exception to this rule.
- **Adjacent Vowels**: In Emakhuwa, the occurrence of adjacent vowels within a word is not grammatically permitted, except for the long vowels "aa", "ee", "ii", "oo", "uu". If any word violates this rule, we assign this feature a value of "1"; otherwise, the value assigned is "0".

Table 3: Emakhuwa's Alphabet

a	aa	c	e	ee	f	h	i
ii	k	kh	l	m	n	ny	ng
o	oo	p	ph	r	s	t	th
tt	u	uu	v	w	x	y	

4.2. Similarity-based features

Similarity-based features were derived using a quantitative approach. We estimate the similarity between word pairs by considering two categories of similarities: lexical and phonetic.

Lexical similarity evaluates the similarity between two sequences from a lexical standpoint. Phonetic similarity, however, assesses how two sequences sound similar. To estimate similarity in both cases, we use the Levenshtein Edit Distance. However, the key distinction lies in their approach: while lexical similarity is calculated directly, for phonetic similarity, we first need to convert the sequences into their phonetic representations. Instead of opting for the common practice of transliterating both loanwords and Emakhuwa words into the International Phonetic Alphabet (IPA) using packages like Epitran (Mortensen et al., 2018), we used the Soundex algorithm (Russell, 1918) to transform the phonetic representations of both loanwords and Emakhuwa words. This choice is made because available tools lack support for Emakhuwa. We used specifically the SoundexBR[2] variant, which is designed for handling Portuguese inputs.

Unlike previous works, we calculate the similarity between the target word (i.e., the word we want to attest) and a potential candidate donor word, which is obtained as explained below. This strategy offers the advantage of requiring only one word as input during training and inference, as opposed to previous works that demanded a pair of donor and recipient words while estimating similarly. This confers resilience when the loanword (i.e., the recipient word) might show spelling discrepancies, a common occurrence within Emakhuwa.

4.2.1. Donor candidate generation

Figure 1 shows the framework of our proposed model, where the candidate donor word is generated in a two-step process:

1. **Translation Candidate**: Given the target word, denoted as w, we use a sequence-to-sequence model trained on a bilingual dictionary of loanwords (see Table 2) that operates at the character level. We only used donors and recipients from the training set. This model takes the Emakhuwa word, w, as input and generates a candidate Portuguese translation, denoted as w'. The assumption is that the model will acquire knowledge based on the patterns observed among these loanword pairs. Therefore, it will learn how to adapt (or "translate") Emakhuwa words back into Portuguese, enabling the generation of words that closely resemble Portuguese vocabulary.
2. **Donor Candidate**: After obtaining the translation candidate, w', we conduct a dictionary lookup to check if it is a valid Portuguese word. If a match is found, we take it as the donor candidate. If not, we perform spell-check the trans-

[2] https://github.com/dmarcelinobr/SoundexBR

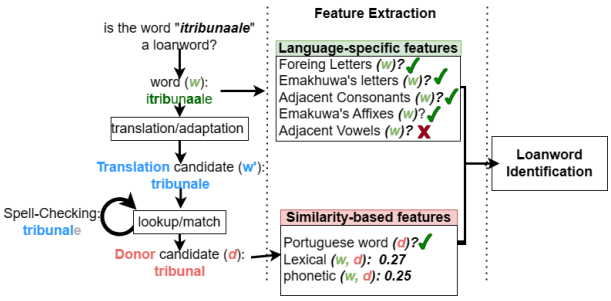

Figure 1: The framework of our proposed model

lation candidate and then conduct a second lookup. If no match is found during the second lookup, the translation is kept as the donor candidate (before spelling correction). For lookup and spelling correction, we use the Portuguese dictionary from the Natura Project[3], which supports European Portuguese.

4.2.2. Sequence-to-Sequence Models models for donor candidate generation

This section explores using different models for character-level sequence-to-sequence Emakhuwa-Portuguese word translation. For that, we trained different models for performance comparison: two are based on a Recurrent Neural Network architecture, and the other is based on the transformer. The difference between the Sequence-to-Sequence models is that one incorporates an attention mechanism as described below.

Sequence-to-Sequence model (seq2seq) : The input to the model consists of a sequence of characters denoted as $X = \langle x_1, ..., x_T \rangle$, and its output is the corresponding translation denoted as $Y = \langle y_1, ..., y_Z \rangle$. Our architecture incorporates two main components: an Encoder and a Decoder (see Figure 2).

- **Encoder** Our Encoder uses a bidirectional Recurrent Neural Network (RNN) with LSTM (Long Short-Term Memory) units. This design enables the Encoder to read the input characters in both directions simultaneously, i.e., from left to right and right to left.
- **Decoder** The Decoder component generates the target translation sequence based on the representation obtained from the Encoder. It uses another LSTM-based RNN to produce the translated output step-by-step. The Decoder considers the previous output and, at each time step determines the next target symbol.

[3] https://natura.di.uminho.pt/wiki/doku.php?id=dicionarios:main

Sequence-to-Sequence model with Attention (seq2seqATT) : This is similar to the Sequence-to-Sequence model except for incorporating an attention mechanism. The attention mechanism (Bahdanau et al., 2015; Vaswani et al., 2017) plays a crucial role in this architecture. It allows the model to focus on relevant parts of the source characters while generating the translation. The attention mechanism dynamically assigns weights to different parts of the input sequence, emphasizing the most informative components during the decoding process.

Transformer (transformer) : The transformer architecture (Vaswani et al., 2017) comprises an encoder and a decoder. The encoder processes the input sequence in parallel, using stacked layers with self-attention and feed-forward sub-layers to generate character representations. The self-attention mechanism weighs character importance, while the feed-forward network captures patterns. The decoder also uses self-attention mechanisms and attends to the encoder's output, generating the output sequence character by character. It relies on the encoded representation of the input sequence for conditioning.
The implementation details for each model is provided below:

- seq2seq: We set the learning rate to 0.001, with a batch size of 64 samples. The encoder and decoder input sizes are determined by the vocabulary sizes of Emakhuwa and Portuguese. Both encoder and decoder employ 300-dimensional embeddings, sharing a hidden size of 1024 to maintain consistency. To prevent overfitting, dropout with a rate of 0.5 is applied to both encoder and decoder components. The early stopping patience is set to 100 epochs.
- seq2seqATT: This model employs the same hyperparameters as seq2seq but includes an attention mechanism with a size of 1024.
- transformer: For the transformer model, we use a learning rate of 3e-4 and a batch size of 32. The source and target vocabulary sizes are aligned with the lengths of Emakhuwa and Portuguese vocabularies. Embedding size is 512, and the model uses 8 attention heads along with 3 encoder and decoder layers. A dropout rate of 0.10 is applied. Similar to other models, the early stopping patience is set at 100 epochs.

For all models, a 10% portion of the training set is allocated for validation. We employ the Adam optimization to compute the cross-entropy loss during training. The model selection is based on early stopping, using loss on the validation set.

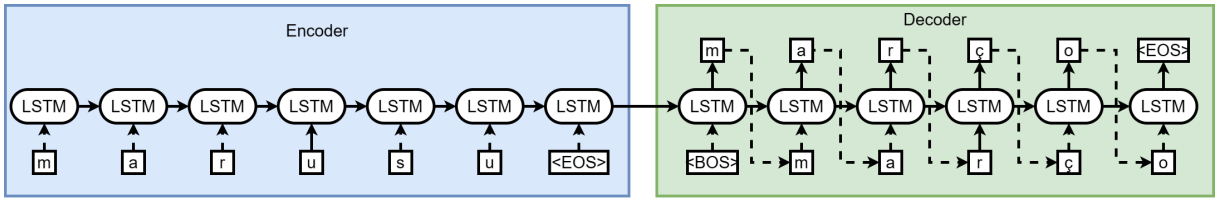

Figure 2: Character-level sequence-to-sequence architecture with LSTM for donor candidate generation.

4.2.3. Evaluation

Using the test set containing donor and recipient words, we employ the Character Error Rate (CER) as a performance metric to assess the effectiveness of the three sequence-to-sequence models. CER quantifies the proportion of incorrect characters in the model's predictions. It is calculated by determining the edit distance between the predicted character sequence and the reference one, then dividing it by the number of characters in the reference.

Table 4 shows each model's performance results. They suggest that the transformer model exhibits a slight advantage over both the "seq2seq" and "seq2seqATT" models in terms of generating correct character sequences, which, in our case, are donor candidates. Thus, we used the "transformer" model to generate donor candidates in all experiments reported in this study.

Table 4: Sequence-to-Sequence models Performance Comparison

Model	CER↓
seq2seq	4.8%
seq2seqATT	4.2%
transformer	4.0%

5. Loanword prediction

This section delves into the performance comparison of various classification algorithms for loanword detection in Emakhuwa. We compare the effectiveness of traditional machine learning models utilizing handcrafted features with the transformer-based model CANINE (Clark et al., 2021).

5.1. Traditional models

We conducted a performance comparison of classification models on our test set, specifically, Logistic Regression (LR), linear Support Vector Machine (SVM), Random Forest (RF), and a Neural Network (NN). For the NN, we followed the same configuration described in (Nath et al., 2022), which includes three layers with hidden units of 512, 256, and 128, respectively, all activated by ReLU. We used 10% dropout and a final sigmoid activation function. Similar to (Nath et al., 2022), it was trained for 5,000 epochs, with early stoping using Adam optimization and binary cross-entropy loss with 20% validation set to prevent overfitting. We used the training and test set described in Section 3.1, in which the class label is either 1 for loanwords or 0 for non-loanwords. For the remaining models, we used the default configuration from scikit-learn (Pedregosa et al., 2011).

5.2. CANINE model

We expanded our analysis to incorporate a pre-trained model for downstream classification tasks. Thus, rather than handcrafting features for our classifier, we leveraged the capabilities of a multilingual pre-trained model. As such, we fine-tuned the CANINE model (Clark et al., 2021), which is a tokenization-free model based on mBERT(Devlin et al., 2019). Specifically, we used the CANINE-C variant, which employs a character-level encoder-only architecture and was pre-trained on a massive dataset encompassing 104 languages, including Portuguese and African languages from the same language family as Emakhuwa. We fine-tuned the pre-trained CANINE model and experimented with two distinct strategies:

- CANINE$_{one}$: we provide a single input corresponding to the target word in Emakhuwa. In simple terms, when given an input in Emakhuwa, the model's task is to classify it as positive or negative.
- CANINE$_{concat}$: In the second strategy, we concatenate the target word with its corresponding match. This concatenation involved placing a special [CLS] token at the beginning of the target input and the special [SEP] token after, then concatenating with the corresponding match from the lookup process explained above. To put it plainly, when given an input in Emakhuwa and its "potential" Portuguese donor, the model's objective is to classify it as either positive or negative.

6. Results and Analysis

This section presents the performance comparison of the models for loanword detection. Table 5 summarizes the results achieved by each model in terms of precision (P), recall (R), and F1-score (F1).

Table 5: Performance results

	Results (%)		
Model	**P**	**R**	**F1**
RF	62	89	73
LR	62	89	73
SVM	62	89	73
NN	61	89	73
CANINE$_{one}$	93	92	92
CANINE$_{concat}$	94	92	93

The findings suggest that the classification models display strong performance in identifying loanwords. All traditional machine learning models achieved an F1-score of 73%. Notably, the model fine-tuned from CANINE significantly outperformed the traditional machine learning models. Where CANINE$_{concat}$ shows a slight performance advantage over CANINE$_{one}$, achieving 94% Precision, 92% Recall, and 93% F1-score.

6.1. Feature Analysis

In this section, we conduct an ablation study to discuss the influence of different features on building traditional machine learning models for loanword detection based on the results presented in Table 6. The features analyzed include language-specific (LS) and similarity (S) features.

Language-specific Features All the traditional machine learning classifiers consistently produced similar or nearly identical outcomes in terms of language-specific features, achieving a Precision of 62%, Recall of 89%, and an F1-score of 73%, when combined. These findings strongly indicate the crucial role of language-specific information in identifying loanwords, particularly when looking at particular letters or affixes, as they boost recall performance by 94%. For precision, only adjacency is demonstrated to help in loan detection. But overall, language-specific features possess sufficient efficacy for detecting loanwords in the context of Emakhuwa.

Similarity Features Considering only similarity features, the models depicted varied performance, with an F1-score ranging between 43% and 46%, significantly lower than language-specific features. Nevertheless, the models consistently performed well when combining language-specific and similarity features. This suggests that combining lexical and similarity features provides complementary information that enhances the traditional machine learning models' ability to distinguish loanwords from non-loanwords.

6.2. Error Analysis

This section presents an error analysis; for that, we only consider the CANINE$_{concat}$, our most promising model from prior experiments.

Table 6: Influence of Features Comparison. LS-Letter Features capturing both foreign letters and those specific to Emakhuwa. LS-Adjacency Features focusing on identifying adjacent vowels and consonants. LS-Affixes Features finding the presence of affixes in the target word.

		Results(%)		
Model	**Features**	**P**	**R**	**F1**
LR	LS-letters	51	94	66
	LS-adjacency	69	29	41
	LS-affixes	51	94	66
	LS (all)	62	89	72
	S	45	71	55
	LS+S	62	89	73
RF	LS-letters	51	94	66
	LS-adjacency	69	29	41
	LS-affixes	51	94	66
	LS-(all)	62	89	73
	S	46	73	56
	LS+S	62	89	73
SVM	LS-letters	51	94	66
	LS-adjacency	69	29	41
	LS-affixes	51	94	66
	LS-(all)	62	89	73
	S	43	86	57
	LS+S	62	89	73
NN	LS-letters	51	94	66
	LS-adjacency	69	29	41
	LS-affixes	51	94	66
	LS-(all)	61	89	73
	S	45	72	55
	LS+S	61	89	73

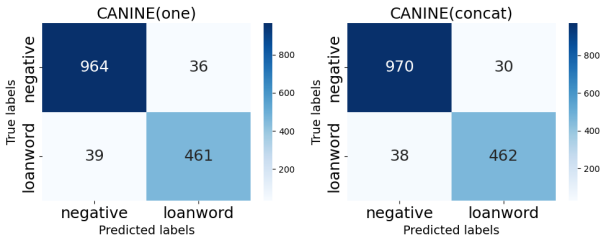

Figure 3: Confusion Matrix, left CANINE$_{one}$ and right CANINE$_{concat}$

6.2.1. False Negatives

We analyzed the error by consolidating a comprehensive list of false negatives generated by the CANINE$_{concat}$ model. Figure 3 shows that our test set revealed 38 false negatives. Upon closer examination, we found that these instances resemble Emakhuwa native words. For instance, the word "mapoweeta", derived from the Portuguese word "poetas". We suspect that the model was misled

Table 7: Results of CANINE model trained on Nath et al.'s dataset

		de-it	ro-fr	en-de	id-nl	kk-ru	hu-de	hi-fa	ca-ar	zh-en	ro-hu	en-fr	fi-sv	az-ar	de-fr	fa-ar	pl-fr
P		100	96	98	99	98	100	97	75	98	99	97	98	98	98	97	97
R	Nath	92	99	99	99	100	93	99	30	93	93	99	98	98	99	97	97
F1	(Nath et al., 2022)	96	98	98	99	99	96	98	43	95	96	98	98	98	98	97	97
P		57	65	66	68	70	65	64	67	68	62	65	68	67	63	60	70
R	CANINE$_{one}$	100	99	88	97	98	100	82	60	59	90	91	93	93	85	96	94
F1		72	79	76	80	81	79	72	63	63	73	76	78	78	73	74	80

by the presence of the radical "weet". Since, in the training set, this radical is often linked to a negative label due to the verb "weetta", which means walk in English. Thus, depending on the context, negation, and tense, this radical can be combined with various affixes, such as "m-weet-teke," "weet-tale", "weet-taka", etc. This led the model to infer that "mapoweeta" is a native Emakhuwa word. This observation applies to all false negatives.

6.2.2. False Positives

Upon examining the outputs of CANINE$_{concat}$, and as we expected, we observed that hard negatives contributed to the presence of false positives. Interestingly, only 16 out of 30 false positives were hard negatives. This indicates that the model exhibited an ability to discern the distinctions between loanwords and native Emakhuwa words despite their similarities.

6.3. Other languages

The CANINE$_{one}$ model demonstrates impressive accuracy in loanword inference from a single input. This approach is preferable and practical for real-world loanword detection scenarios, where, typically, the task is to identify whether a word was borrowed from another language. To explore the model's generalizability across languages, we fine-tuned the dataset from Nath et al. 2022 and applied CANINE$_{one}$ to 16 language pairs. These pairs represent a spectrum of donor-recipient relationships, including: German-Italian (de-it), Romanian-French (ro-fr), English-German (en-de), Indonesian-Dutch (id-nl), English-French (en-fr), Kazakh-Russian (kk-ru), Hungarian-German (hu-de), Hindi-Farsi (hi-fa), Catalan-Arabic (ca-ar), Chinese-English (zh-en), Romanian-Hungarian (ro-hu), Finnish-Swedish (fi-sv), Azerbaijani-Arabic (az-ar), German-French (de-fr), Farsi-Arabic (fa-ar), and Polish-French (pl-fr).

Table 7 presents the results, which we also compared with results for the study of Nath et al.. All models fell short in terms of F1 score except for the ca-ar language pair. Interestingly, the models exhibited competitive recall results but struggled with precision. Among the recipient languages, kk-ru, id-nl, and pl-fr scored top F1, achieving 81%, 80%, and 80%, respectively. Notably, all agglutinative languages (i.e., Kazakh, Hungarian, and Indonesian) exhibited strong performance, exceeding 78% F1. Our interpretation of these results is that agglutinative languages provide rich linguistic cues that enable the model to learn how to discern loanwords even without prior knowledge of their donor counterparts.

7. Conclusion

In conclusion, this research introduces an innovative method for automating the detection of loanwords in Emakhuwa. Our approach leverages language-specific and similarity features to effectively identify words loaned from Portuguese. Also, fine-tuning the CANINE models. This is a significant contribution, as loanword detection is extremely promising for enhancing various NLP tasks. Beyond the method, we also provide a dataset associated with loanword detection in Emakhuwa. These datasets serve as valuable resources for researchers and practitioners in NLP in low-resource languages like Emakhuwa. Furthermore, the proposed method can be applied to other language pairs with similar characteristics, with promising effects for potential application to language pairs in which one language benefits from enriching its vocabulary through borrowing from another language.

8. Acknowledgements

This work was financially supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC) as well as supported by the Base (UIDB/00022/2020) and Programmatic (UIDP/00022/2020) projects of the Centre for Linguistics of the University of Porto. Felermino Ali is supported by a PhD studentship (with reference SFRH/BD/151435/2021), funded by Fundação para a Ciência e a Tecnologia (FCT).

9. References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Felermino D. M. A. Ali, Andrew Caines, and Jaimito L. A. Malavi. 2021. Towards a parallel corpus of

portuguese and the bantu language emakhuwa of mozambique.

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

BSI. 1973a. *Natural Fibre Twines*, 3rd edition. British Standards Institution, London. BS 2570.

BSI. 1973b. Natural fibre twines. BS 2570, British Standards Institution, London. 3rd. edn.

A. Castor and L. E. Pollux. 1992. The use of user modelling to guide inference and learning. *Applied Intelligence*, 2(1):37–53.

J.L. Chercheur. 1994. *Case-Based Reasoning*, 2nd edition. Morgan Kaufman Publishers, San Mateo, CA.

N. Chomsky. 1973. Conditions on transformations. In *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2021. CANINE: pre-training an efficient tokenization-free encoder for language representation. *CoRR*, abs/2103.06874.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Umberto Eco. 1990. *The Limits of Interpretation*. Indian University Press.

Paul Gerhard Hoel. 1971a. *Elementary Statistics*, 3rd edition. Wiley series in probability and mathematical statistics. Wiley, New York, Chichester. ISBN 0 471 40300.

Paul Gerhard Hoel. 1971b. *Elementary Statistics*, 3rd edition, Wiley series in probability and mathematical statistics, pages 19–33. Wiley, New York, Chichester. ISBN 0 471 40300.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Yoonjung Kang. 2011. Loanword phonology. *The Blackwell companion to phonology*, pages 1–25.

V I. Levenshtein. 1965. Dvoiˇcnye kody s ispravleniem vypadenij, vstavok i zameš ˇcenij simvolov[binary codes with correction of deletions], insertions, and replacements. *Doklady Akademij Nauk SSSR*, 163(4)(8):845–848.

Johann-Mattis List and Robert Forkel. 2022a. Automated identification of borrowings in multilingual wordlists. *Open Research Europe*, 1:79.

Johann-Mattis List and Robert Forkel. 2022b. Automated identification of borrowings in multilingual wordlists. *Open Res Europe*, 1:79.

Chenggang Mi. 2023. Loanword identification based on web resources: A case study on wikipedia. *Computer Speech & Language*, 81:101517.

Chenggang Mi, Lei Xie, and Yanning Zhang. 2020. Loanword identification in low-resource languages with minimal supervision. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3):1–22.

Chenggang Mi, Yating Yang, Lei Wang, Xiao Li, and Kamali Dalielihan. 2014. Detection of loan words in uyghur texts. In *Natural Language Processing and Chinese Computing*, pages 103–112, Berlin, Heidelberg. Springer Berlin Heidelberg.

Chenggang Mi, Yating Yang, Lei Wang, Xi Zhou, and Tonghai Jiang. 2018. Toward better loanword identification in Uyghur using cross-lingual word embeddings. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3027–3037, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Chenggang Mi, Shaolin Zhu, and Rui Nie. 2021. Improving loanword identification in low-resource language with data augmentation and multiple feature fusion. *Computational Intelligence and Neuroscience*, 2021.

John Miller, Emanuel Pariasca, and Cesar Beltran Castañon. 2021. Neural borrowing detection with monolingual lexical models. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 109–117, Online. INCOMA Ltd.

John E. Miller and Johann-Mattis List. 2023. Detecting lexical borrowings from dominant languages in multilingual wordlists. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2599–2605, Dubrovnik, Croatia. Association for Computational Linguistics.

David R. Mortensen, Siddharth Dalmia, and Patrick Littell. 2018. Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori S. Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484. ACL.

Abhijnan Nath, Sina Mahdipour Saravani, Ibrahim Khebour, Sheikh Mannan, Zihui Li, and Nikhil Krishnaswamy. 2022. A generalized method for automated multilingual loanword detection. In *International Conference on Computational Linguistics*.

John E Ortega, Richard Alexander Castro Mamani, and Jaime Rafael Montoya Samame. 2021. Love thy neighbor: combining two neighboring low-resource languages for translation. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 44–51.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

R. Russell. 1918. United states patent 1,261,167.

Charles Joseph Singer, E. J. Holmyard, and A. R. Hall, editors. 1954–58. *A history of technology*. Oxford University Press, London. 5 vol.

Jannik Strötgen and Michael Gertz. 2012. Temporal tagging on different domains: Challenges, strategies, and gold standards. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3746–3753, Istanbul, Turkey. European Language Resource Association (ELRA).

S. Superman, B. Batman, C. Catwoman, and S. Spiderman. 2000. *Superheroes experiences with books*, 20th edition. The Phantom Editors Associates, Gotham City.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.