Automatic Detection of Borrowings in Low-Resource Languages of the Caucasus: Andic branch

Konstantin Zaitsev **Anzhelika Minchenko**

HSE University

knzaytsev@edu.hse.ru aminchenko@edu.hse.ru

Abstract

Linguistic borrowings occur in all languages. Andic languages of the Caucasus have borrowings from different donor-languages like Russian, Arabic, Persian. To automatically detect these borrowings, we propose a logistic regression model. The model was trained on the dataset which contains words in IPA from dictionaries of Andic languages. To improve model's quality, we compared TfIdf and Count vectorizers and chose the second one. Besides, we added new features to the model. They were extracted using analysis of vectorizer features and using a language model. The model was evaluated by classification quality metrics (precision, recall and F1-score). The best average F1-score of all languages for words in IPA was about 0.78. Experiments showed that our model reaches good results not only with words in IPA but also with words in Cyrillic.

1 Introduction

Field linguistics develops and practises methods for obtaining information about a language unknown (or little known) to the researcher based on work with native speakers. Such languages are called low-resource languages; they represent a group of languages for which the development of information technology is insufficient. There are a number of criteria (for example, speech processing, speech recognition, automatic translation, and others) according to which experts classify specific languages as low-resource.

Lexical borrowings are very common to languages, including those with few resources; this phenomenon is caused by interlingual interaction and influence. If borrowings from languages with limited resources (for example, Botlikh) are effectively identified, then automatic detection of borrowings with a universal base for related languages can be created and used. This article studies the method of identifying borrowings in low-resource Andic languages on a linguistic basis. It implies that the model imitates the borrowing rules in the receiving language based on identifying the most relevant n-grams and generating words based on the identified borrowing patterns.

Many tools for working with Andic languages are currently being developed, such as morphological parsers. Even though each language is unique and has linguistic properties, all of them are underprivileged and endangered, as the number of their speakers is constantly decreasing, and transmission from generation to generation becomes unstable. That makes developing any NLP tools essential as it can help in their further exploration and potential revival. In addition, the detection of borrowings will help to study the language more deeply and try to preserve its identity. In the future, the work could be used to create a universal transliteration so that as many linguists as possible could work with languages and, for example, with texts.

The paper's main goal is to explore the possibility of automatic borrowing detection without the usage of a bilingual dictionary since automation can contribute to future field studies of target languages. The limited amount of available data complicates the situation by reducing the number of possible analysis methods that can be implemented. The first task was to analyze the existing dictionaries. The analysis showed that the dictionaries had duplicates, which were later removed. After removing duplicates, the general borrowing rules were determined and a baseline was written with further verification of its quality. The next step was to calculate and describe insights that helped to improve the quality of the baseline. As a result, previous steps helped to cope with implementing a language model for generating additional features. To assess the quality, it was

Proceedings of the first workshop on field linguistics, pages 34–41
October 16, 2022.

necessary to perform tasks such as writing a quality metric for the language model and statistical analysis of features. These steps will be discussed in more detail in the following sections.

The rest of the article is structured as follows: the second section briefly overviews the target languages and their problems. After that, the review of the relevant literature in computational linguistics continues. The third section describes the methodology and strategies that have been used to implement the structures of each language. The fourth and fifth sections evaluate and discuss the results obtained from the available language data. The conclusion also discusses the problems that have occurred in working on the model, as well as a short description of plans for the future.

2 Literature review

2.1 Low resource language

The term "low-resource languages" (or under-resourced languages) was initially proposed by the Dutch scientist S. Krauwer. This concept refers to natural languages with some (or all) of the following properties (Vincent, B., 2004):

- the lack of their writing system or stable spelling;
- lack of qualified linguists and translators for the given language;
- limited distribution on the Internet;
- lack of electronic resources for language and speech processing, including monolingual corpora, bilingual electronic dictionaries, spelling and phonetic transcriptions of speech, pronunciation dictionaries, and more.

2.2 Theories of borrowing analysis

The term "borrowing" refers to complete language change, a diachronic process that once began as an individual innovation but then spread throughout the speech community. The most common borrowing theories for under-resourced languages are based on language rules or systems based on the constraints of those rules. While a constraint-based system basically ends up within optimality theory, rules describe how adaptation occurs and is set according to a particular borrowed word in the language's phonology (Jacobs, H., & Gussenhoven, C., 2000). Therefore, rules must be added for each specific borrowing, considering the functional aspect of speech. In addition, the rule-based model only includes rules for a particular language, so each language needs either a separate word adaptation system or a family-wide one.

A constraint-based system is analogous to a rule-based system. Constraints are included in the Optimality Theory (OT) structure. Basically, all studies of borrowings are based on this system (Turchin, P., 2010). In a constraint-based system, several constraints are defined and ranked. The input of the model is the source word with its pronunciation in the source language.

As for research in the field of borrowings by computer linguists, there are several main approaches. They can be based on both neural networks and the Optimality Theory. Neural networks are used to determine loanwords in the Uyghur language (low resource) in (Mi et al., 2018). The authors used a recurrent neural network with BiLSTM architecture, training it on a dataset with borrowings in the Uighur language. As a result, the model showed promising results, as presented in Table 1 (Mi et al., 2018). “Chn”, “rus” and “arab” suffixes mean Chinese, Russian and Arabic languages respectively.

For lexical borrowings, OT is also used. The usage of OT is described in (Tsvetkov, Y., & Dyer, C., 2016). Authors' implemented model was based on OT, and it used various restrictions for Swahili, which contains borrowings from Arabic (Table 2). Similar restrictions the model uses allow one to get better results compared to simple implementations of borrowing detection.

As for neural network approaches, a possible problem is a lack of sufficient data and the need for

Model	Pchn	Rchn	F1chn	Prus	Rrus	F1rus	Parab	Rarab	F1arab
CRFs	69.78	62.33	66.35	71.64	63.25	67.18	72.50	65.32	68.72
SSIM	66.32	77.28	71.38	75.39	70.02	72.61	73.76	67.51	70.50
CIBM	78.82	68.30	73.18	81.03	73.22	76.93	75.22	70.71	72.90
RNN	78.97	79.20	79.08	82.55	75.93	79.10	83.26	77.58	80.32
Ours	**80.24**	**81.02**	**80.63**	**82.95**	**76.30**	**79.49**	**84.09**	**78.28**	**81.08**

Table 1. Experimental results of borrowings identification models based on a recurrent neural network with BiLSTM architecture.

/ ɛg/	DEP-IO	MAX-IO	ONSET	NO-CODA
a. ☞ ɛg			*	*
b. ɛgɘ	*!		*	
c. ɛ		*!	*	
d. ʔɛg	*!			*

Table 2. Restrictions for Swahili in the study by Tsvetkov and Dyer.

enormous computing power. In addition, OT has the disadvantage of building restriction systems for each Andic language. Such an approach will not have universality property, and its implementation will take a long time. For these reasons, we have chosen a baseline based on logistic regression, which will be presented later in the paper.

2.3 Materials for research

The collection of Andic language dictionaries (Moroz, G. et al., 2021) is used as a dataset. In total, at the moment, it contains nine (9) languages; however, for our study, we analyze only eight (8) of them since there is not enough data for the Tokita for a full-fledged study. The dataset contains two Botlikh dictionaries used in the work as sources for one language, without separation. Table 3 shows the glottocode of the language (a bibliographic database of obscure languages), its name, and the number of words in it.

Glottocode	Language	Number of Words
akhv1239	Akhvakh	14007
andi1255	Andi	6144
bagv1239	Bagvalal	12706
botl1242	Botlikh	21483
cham1309	Chamalal	9721
ghod1238	Godoberi	7423
kara1474	Karata	6650
tind1238	Tindi	12419

Table 3. Glottocode of low-resource Andic languages.

Each word in the database contains a form translated into the International Phonetic Alphabet (IPA), its canonical form (lemma), and an indication of whether the word is borrowed or not (bor). In turn, each borrowing has a short description, indicating the language from which the word came (borrowing_source_language). Some words can have different meanings or borrowing source languages. To make the task easier, we dropped duplicates and kept last occurrence of the dropped word. This approach is not quite accurate, but the number of such cases is very low. Column “meaning_ru” is written in Russian but for this paper it has an English translation. All data was collected by authors of the dataset, so we did not make any transliteration, normalization and so on. An example of a dataset with important columns for the model is presented in Table 4.

3 Method

3.1 Baseline training

The dataset presented in the previous section is at the heart of our research into language patterns and baseline learning. Since the task is to determine borrowing, models for classification are suitable for this. Also, words in IPA will be used to train the model, as they give a cleaner characteristic of borrowing. In addition, most of the work is done in the IPA, as it, unlike transcription in Cyrillic, marks the sounds of the language, which helps to conduct a cleaner analysis.

Of all classifier models, logistic regression was chosen. We decided to use TfIdf Vectorizer to transform list of words in IPA to matrix with tf-idf weights. In this matrix rows are input words and columns are symbol n-grams of each input word. To work correctly with these words, we wrote the specific token pattern that removes hyphens and splits word to IPA-symbols. In addition, we added from 2 to 4 n-grams to n-grams hyperparameter of the vectorizer. The resulting combination of models was trained in each dataset language. Training took place on the training set, validation on the test set,

lemma	ipa	glottocode	bor	borrowing_source_language	meaning_ru
аба'далIи	a-b-'a-d-a-t-łː-i	akhv1239	1	arab	Eternal
а/б/а'жве	a-b-'a-ʒʷ-e	akhv1239	0	NaN	everlasting
а/б/ажу'рулъIа	a-b-a-ʒ-'u-r-u-t-ł-a	akhv1239	0	NaN	communicate

Table 4. Dictionary description for the Akhvakh language.

Language	Precision	Recall	F1
Ahvakh	0.90	0.57	0.60
Andi	0.80	0.56	0.58
Bagvalal	0.81	0.60	0.63
Botlikh	0.88	0.74	0.78
Chamalal	0.97	0.51	0.50
Godoberi	0.89	0.61	0.65
Karata	0.96	0.51	0.49
Tindi	0.97	0.53	0.54

Table 5. Metrics for Andic languages obtained after training the baseline.

while the partition was based on the 80/20 principle. The macro average f1-score metric was used to assess the model's quality since the classes in the dataset are not balanced. After training and testing the models, it turned out that their quality was low. It was easier for Baseline to say that a word was not borrowing than the other way around. The metrics for this model for each language can be seen in Table 5.

3.2 Selection of hyperparameters

Since the baseline quality turned out to be poor, the next step was to select hyperparameters using heuristics for the vectorization model. We decided to use CountVectorizer instead of TfIdfVectorizer. This decision was based on several experiments with the same hyperparameters. CountVectorizer works like TfIdfVectorizer except for output. The output of CountVectorizer is the matrix of counted words. We added hyperparameters (min_df = 0.001, max_df = 0.1) responsible for filtering rare and frequent n-grams to get rid of noise. The number of features limitation was also removed. Experiments showed that chosen hyperparameter values are the most optimal for the model.

This implementation of the vectorization model significantly increased the model's quality, but in some languages, the F1-score remained low. To fix this problem, we analyzed the n-grams (or features) from the vectorizer matrix. The analysis showed that some of the features contribute the most to the model's quality. From these features we selected some of them which value corresponds to the set hyperparameters. Then we filtered part of selected features by a threshold value. It allowed us to select features more like the borrowing patterns we studied in languages. For each word in the dataset, it was determined whether n-grams are included in this list of features. We added a positive coefficient for the word in the case of such a feature in the n-gram of the word. The optimal coefficients and hyperparameters were selected by experiments. As a result, this approach allowed us to improve the model by small values. In the next sections this model is called as BF (baseline with features). Table 6 presents the quality metrics for the model.

3.3 Language model approach

Borrowings are characterized by the fact that they may contain those phonemes that are not typical for the receiving language, which belongs to OOV (out of vocabulary). Accordingly, such sequences may indicate that the word is borrowed. This knowledge underlies the model built on the language model on Markov chains (on n-grams), which was implemented at the next stage of the study. For the language model, a perplexity metric (Jurafsky, D., & Martin, J., 2009) was also developed to evaluate the similarities of a word to a language.

Since perplexity shows how unfamiliar the word is for the model, it can be said that the model

Language	Precision	Recall	F1
Ahvakh	0.79	0.72	0.74
Andi	0.75	0.69	0.71
Bagvalal	0.80	0.71	0.74
Botlikh	0.86	0.83	0.85
Chamalal	0.80	0.65	0.70
Godoberi	0.82	0.77	0.79
Karata	0.76	0.65	0.69
Tindi	0.73	0.65	0.68

Table 6. Metrics for Andic languages obtained after selecting hyperparameters.

trained on the language will have a lower perplexity value for non-borrowings than for borrowings. For verification, an auxiliary dataset was collected, consisting of the perplexities of each word. The language model was trained for each language of the initial dataset. The model calculated the perplexity value for the input word over several n-grams. After the calculation, the value was written to the dataset, which consisted of a word in the IPA, a lemma, a borrowing label, and perplexity values for each n-gram.

When splitting the dataset by language, we got results that visually confirmed the hypothesis. Moreover, the Wilcoxon-Mann-Whitney nonparametric statistical test confirmed the hypothesis about high perplexity of borrowing words put forward; at the same time, it can be seen that the differences in perplexities are most pronounced for trigrams. Visualization is shown in Figure 1.

The difference between perplexities further helped to implement a model that, according to trigrams, speaks of borrowing. In our study, we conducted experiments that showed that trigrams work better than bigrams (four-grams were not considered due to the identical distributions). Thus, trigrams were chosen because they best represent foreign words and experiments with bigrams and trigrams. The model is based on a language model that works like those presented above. The difference is that the language model is trained on non-borrowings since borrowings are characterized by combinations of phonemes that may not be in the language.

The language model helps to get new features from words using the algorithm. Each input word is divided into trigrams, checked in the language model: if it does not have such a trigram, then the word is borrowed and is set some positive coefficient that was selected by experiments. Otherwise, the highlighted word has a negative rate. With the help of that algorithm, a list of borrowing marks was collected for each word and added to other features.

3.4 Combining Models

Implicit knowledge of phoneme sequences can improve a regression model, as it can sometimes generate false positives on its own. For example, suppose some algorithm generates a word produced by a language model trained on borrowings. In that case, it may be borrowing since it contains a sequence of phonemes that are not in the language, although the opposite was meant. Alternatively, there may be such a situation when the language model does not have many examples. In this case, the probability of error also increases. For these reasons, additional knowledge about the language (in this case, the use of regression) can improve the results.

To implement such a model, we combined the results of the regression and trained language model. As a result, the model began to work better, although, in some languages, the quality decreased

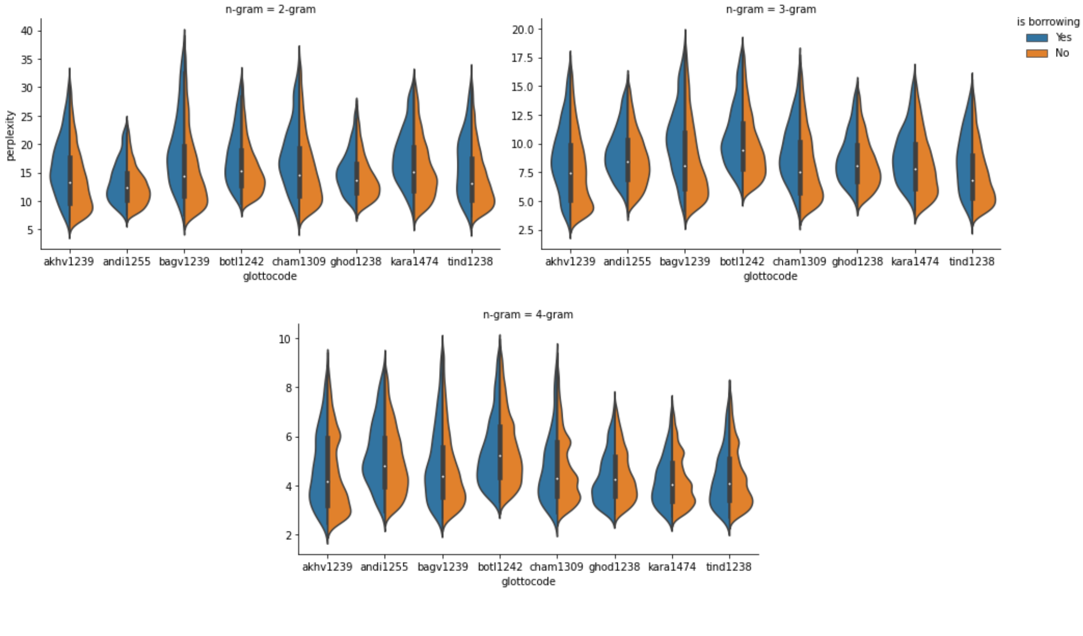

Figure 1. Graphs of the obtained results of perplexity.

Language	Precision	Recall	F1
Ahvakh	0.75	0.83	0.78
Andi	0.72	0.76	0.74
Bagvalal	0.78	0.82	0.80
Botlikh	0.80	0.88	0.83
Chamalal	0.76	0.80	0.78
Godoberi	0.78	0.86	0.81
Karata	0.70	0.73	0.71
Tindi	0.70	0.77	0.73

Table 7. Cross-validation results.

slightly compared to the previous model. The model was tested on test split, but the hyperparameters were fitted by K-fold validation, which showed high quality. The cross-validation results can be seen in Table 7.

4 Results

In addition to learning words in the IPA, the model was also trained on lemmas (BFLMlem). This experiment was carried out to compare the purity of words written in phonemes and graphemes. As a result, it turned out that the quality of the model is higher than BFLM (baseline with selected features and the language model) on IPA. Hence, the BFLM will work well for words written in IPA and Cyrillic both. A comparison of the models implemented in the article, according to the F1-score metric, is presented in Table 8.

We compared our models with others mentioned in related works. We calculate mean precision, recall, and F1-score metrics from our experiments and the results in other articles. Comparison shows that our models work slightly worse than the others, but scores remain high. Hence, simple models with feature extraction based on linguistics knowledge, such as knowing about OOV, can show results close to complicated neural network architecture models. Models' comparison is presented in Table 9.

In addition to the experiments, we tested BFLMlem on random letters and numbers. We got 0.93 mean accuracy of language models. Besides, we examined BFLM trained on IPA on English words and got 0.51 mean accuracy.

5 Discussion

In continuation of the idea of assessing perplexity in words, neural network models can be used in the future. A recurrent neural network is perfect for this. The neural network can be trained on borrowings and then generate new words and find specific patterns.

The dictionary does not fully reflect the quality of the model since it does not consider various morphological features, such as declension. For this reason, the model must be tested on work with texts. This way, it will be possible to take each word in context and determine whether it is borrowing. On the other hand, texts in languages are not presented in IPA but are written in Cyrillic. In this case, it will be possible to use the epitran tool, having previously written the rules for converting graphemes to phonemes (Mortensen, R. D., Dalmia, S., & Littell, P., 2018). In addition to the problem with the transformation, there is also the possibility that word declensions will also negatively affect the model. In general, this approach will show the actual quality of the model and can further help field linguists.

Now the model works for each language, classifying the words in it as borrowing. In the future, it may be worth refining the model, adding to it not only a binary classification but also a definition of the language from which the borrowing occurred. In this case, the problem can

Model	Akhvakh	Andi	Bagvalal	Botlikh	Chamalal	Godoberi	Karata	Tindi
Baseline	0.60	0.59	0.63	0.78	0.50	0.65	0.50	0.54
BF	0.73	0.69	0.74	0.84	0.68	0.78	0.68	0.66
BFLM	0.78	0.74	0.80	0.83	0.78	0.81	0.71	0.73
BFLMlem	**0.82**	**0.77**	**0.84**	**0.86**	**0.79**	**0.84**	**0.75**	**0.75**

Table 8. Model quality comparisons.

Model	Precision	Recall	F1
our BFLM	0.75	0.81	0.77
our BFLMlem	0.78	**0.83**	0.80
Neural Network for Uyghur	0.82	0.79	0.80
BiLSTM-CRF for Spanish	**0.91**	0.79	**0.84**

Table 9. Our models' results compare to other research.

be reformulated not within the framework of the classification but within the framework of BIO-encoding, which has already been solved for the Spanish corpus in (Alvarez-Mellado, E., & Lignos, C., 2022). Also, if we consider borrowings separately by language, it makes sense to look at the n-grams characteristic of borrowings from a particular language. Perhaps a combination of such phonemes will also speak of the source language.

In this paper, we proposed methods that can be used in a borrowings detection task. It is possible that our findings might be implemented in other models which find borrowings in low-resource languages. Besides, detected borrowings by the model might be helpful for field linguists working with Andic languages to understand deeply these languages.

6 Conclusion

This article has shown how to solve the problem of classifying borrowings in Andic low-resource languages. For this, a baseline was first used, consisting of logistic regression and TfIdf of the vectorization model. Due to unsatisfactory results, the vectorization model was changed from TfIdfVectorizer to CountVectorizer, and hyperparameters were selected for it. In addition, a simple model based on implicit language knowledge was written. After combining these models, the quality has improved significantly. As a result, our models have scores close to neural network solutions. Hence, simple binary classification can be used in tasks such as detecting borrowings. However, since the model solves a binary classification problem, it cannot tell the origin of the borrowing. In the future, it is planned to supplement the model by teaching it to solve either the problem of multiclass classification or BIO-encoding. For these problems, the future models can be based on the already implemented. Code and research are available on the GitHub repository[1].

References

Alvarez-Mellado, E., & Lignos, C., 2022. Detecting Unassimilated Borrowings in Spanish: An Annotated Corpus and Approaches to Modeling. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 1*, 3868–3888.

Haspelmath, M., 2008. Loanword typology: Steps toward a systematic crosslinguistic study of lexical borrowability.

Jacobs, H., & Gussenhoven, C., 2000. Understanding phonology. *Language, 76*(1), 209. https://doi.org/10.2307/417430

List, J., Moran, S., & Prokić, J., 2013. Automatic detection of borrowings in lexicostatistic datasets.: A workflow for automatic linguistic reconstruction. *Workshop on Quantitative Approaches to Areal Typology.*

Mi et al., 2018. A Neural Network Based Model for Loanword Identification in Uyghur. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Moroz, G. et al., 2021. Comparative Andic dictionary database, v. 0.5 Moscow: Linguistic Convergence Laboratory, HSE University. Available from: https://github.com/phon-dicts-project/comparative_andic_dictionary_database. DOI: 10.5281/zenodo.4782876

Jurafsky, D., & Martin, J., 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Mortensen, R. D., Dalmia, S., & Littell, P., 2018. Epitran: Precision G2P for Many Languages. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Tsvetkov, Y., & Dyer, C., 2016. Cross-lingual bridges with models of lexical borrowing. *Journal of Artificial Intelligence Research, 55*, 63–93. https://doi.org/10.1613/jair.4786

Turchin, P., 2010. Analyzing genetic connections between languages by matching consonant classes.

[1] https://github.com/Knzaytsev/Borrow-Detection

Vincent, B., 2004. Methods to computerize "little equipped" languages and groups of languages.