Language Identification of Guadeloupean Creole

William Soto

▶ **To cite this version:**

William Soto. Language Identification of Guadeloupean Creole. 2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Dec 2020, Montrouge (virtuel), France. pp.54-59. hal-03047144

HAL Id: hal-03047144
https://hal.science/hal-03047144

Submitted on 3 Jan 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Language Identification of Guadeloupean Creole

William Soto
Université de Lorraine, Nancy, France
williamsotomartinez@gmail.com

RÉSUMÉ

L'identification automatique de la langue est une étape de pré-traitement particulièrement utile lorsque l'on traite des données provenant de sources multilingues. Cette étape permet notamment de filtrer les textes en fonction de la langue utilisée et d'appliquer des traitements adéquats. Les langues peu dotées ne sont malheureusement pas toujours supportées par les outils d'identification automatique. Dans cet article, nous présentons un outil d'identification automatique du créole guadeloupéen, qui repose sur une approche à base d'apprentissage automatique pour résoudre ce manque dans une certaine mesure. L'évaluation de notre modèle montre une précision de 89,96% sur l'identification de phrases en créole guadeloupéen provenant de différentes sources, et une précision de 91,04% sur l'identification de phrases dans un corpus de 103 langues.

ABSTRACT

Language Identification of Guadeloupean Creole

Language Identification is a useful preprocessing step that allows filtering and processing information on the best way possible, Improving the efficiency of Language Processing tasks. Under-resourced language, however, are often left out of most off-the-shelf applications for this task. In this article we present the Guadeloupean Creole Language Identification Tool, a Machine Learning (ML) approach to solve the lack of such applications for this under-resourced language. The evaluation of our model shows an 89.96% accuracy when classifying Guadeloupean Creole sentences from different sources and a 91.04% precision on the language when classifying sentences from 103 different languages.

MOTS-CLÉS : Identification Automatique de Langue, Créole Guadeloupéen, Apprentissage Automatique.

KEYWORDS: Language Identification, Guadeloupean Creole, Machine Learning.

1 Introduction

Language identification (LI) consist on determining in which language a given text has been written. As Jauhiainen et al. (2019) point out, Natural Language Processing (NLP) and Information Retrieval (IR) techniques generally assume that all documents given to a system are written in the same language. Because of this, being able to correctly identify and filter out documents that do not math the system's languages is a useful preprocessing step.

When you are guaranteed to have a variety of languages as input to your system, like with social network resources, this step becomes even more relevant. Cases like the SURICATE-Nat[1] project,

[1] See http://www.suricatenat.fr/Suricate-Nat/ for more information

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 53–58, *©2020 CNRS* Attribution 4.0 International.

that collects and analyzes Twitter messages to facilitate information exchange from the ground during natural disasters, is a clear example of this. However, the LI task may not be as efficient in this kind of real life situations since most off-the-shelf tools do not support under-resourced languages, which in turn reduces the extent to which certain communities benefit from these projects.

Creole languages (languages that develop from the mix and simplification of other languages) are a good example of languages left out of most modern LI tools. Although these languages have been studied by linguists for centuries, many of them remain under-resourced when it comes to modern NLP applications. Their low number of speakers and a the lack of standardization are, no doubt, part of the reason behind that.

For this article we focused on one specific case that is usually not present on LI software: Guadeloupean Creole (GC, or Gwadloupéyen). This language developed from XVIII century French and a variety of West African languages (Delumeau, 2006) and accounts for around 600 000 speakers, from which close to 400 000 live in Guadeloupe (Colot and Ludwig, 2013) (a French archipelago in the Caribbean). For the most part it has been a spoken language and lacks a strong spelling standard, although there have been various attempts to change this (see Ludwig et al. (1990); Hazaël-Massieux (1993); Bernabé (2001) about the GEREC orthography). However, as Delumeau (2006) points out the language has become increasingly popular on Guadeloupe.

2 Related Work

Some LI off-the-shelf tools like LangDetect (Nakatani, 2010) and Compact Language Detector 2 (Sites, 2013) rely on probabilistic approaches, most specifically Naive Bayes classifiers. Some other methods, like Whatlang (Brown, 2013), rely on vector-space models. More recent methods are based on FastText (Armand et al., 2016) applying n-gram word embeddings and linear regression, like FastText's own language detection module (Joulin et al., 2016) and Whatthelang (Sangeeth, 2017).

All these methods have proven to be useful to solve LI tasks, but none of them provides support for Guadeloupean Creole. As we pointed out in the Introduction, under-resourced languages are usually excluded from modern NLP applications partially because of the lack of training data. For the case of GC, Millour and Fort (2018) mention some of the the existing resources of GC and detail the state of the art regarding GC corpora. However, the existing source are rather scarce and, to our knowledge, have not been applied to modern ML tasks.

3 Methodology

Following the most recent approaches we decided to use FastText as the base for our classifier. Not only do FastText-based models perform very well, but they are easy to train and to deploy. Our model consists of a FastText supervised classifier, which builds text representations by averaging n-grams and then performs multiple logistics regressions over this text representations. This model allows the use of subword features, which applies the n-grams at the character level and makes it possible to get information about the structure of words as well as supporting out of vocabulary (oov) words.

4 Dataset

As a baseline we decided to use the Tatoeba sentence dataset which contains more than 8 million sentences of 355 different languages. Most notably it, has 2 080 GC sentences. To reduce the training time and the noise added by languages without many samples, we limited the dataset to those languages with at least 1000 sentences, which left us with a total of 103 languages. Table 1 shows the distribution of the most common and uncommon languages on the dataset as well as that of GC.

Most common languages	Samples	Less common languages	Samples
English	1 319 616	Kapampangan	1 475
Russian	759 878	Cebuano	1 472
Italian	702 267	Ottoman Turkish	1 407
Turkish	685 782	Albanian	1 351
Esperanto	618 098	Picard	1 339
German	502 445	Khasi	1 320
French	425 191	Old East Slavic	1 307
Portuguese	358 570	Guarani	1 251
Spanish	317 954	Welsh	1 237
Hungarian	281 093	Slovenian	1 046

Table 1: Number of samples for the 10 most and less common languages in the tatoeba dataset.

To improve the models' ability to identify GC, we enriched the dataset with samples of the language from a diversity of sources. We used a set of transcriptions of spoken Guadeloupean Creole (Glaude, 2013), Caterina Bonan's Corpus of Guadeloupean Creole 2018, the 2012 Simenn Kréyòl collection of texts by the Academié de la Guadeloupe, the lyrics of two Guadeloupean Songs (the LKP song "gwada sé tan nou" and the Akiyo song "Jilo") and articles from the GC Wikipedia, which increased the number of GC sample sentences from 2080 to 4894.

All the datasets were preprocessed by lowercasing all the text and removing punctuation forms in each sentence. Then they were split in training and testing sets, with a 90% of the sentences going to the training set and the other 10% going to the testing set. Table 2 shows the number of GC samples present on each dataset.

Source	Training	Testing	Total
Tatoeba	1872	208	2080
Transcriptions	1354	150	1504
Caterina Bonan	689	78	767
Simmen Kréyòl	327	36	363
Chansons	100	11	111
Wikipedia	63	6	69
Total	4405	489	4894

Table 2: Number of Guadeloupean Creole samples on each dataset.

5 Experiment and Results

When instantiating the FastText supervised classifier we tried with 3 different sizes for the word vectors (16,32 and 64 dimensions) and enabled the use of subword features by setting up the char n-grams to a minimum size of 2 and a maximum size of 4. To set the other parameters of the model we tried the autotune method included in FastText, but after some tests we saw no significant improvement over the default settings, so we kept the default values on all the other parameters with the exception of the number of training epochs which showed an improvement when we doubled it from 5 to 10 epochs.

As mentioned above, the baseline was trained exclusively on the Tatoeba training set and another model was trained on the enriched GC dataset. To evaluate them, two tasks were defined: first we measured the accuracy of the models when presented only with the GC samples from each of the different sources, then we calculated the precision, recall and F1 score of the models when presented with samples from all the languages of the enriched dataset.

Tables 3 and 4 shows the result of the first and second task respectively. When classifying just GC sentences, as expected, the baseline performs well with the Tatoeba examples but accuracy drops with the other sources. In turn, the enriched model performs very well on all the sources. For the second task, when classifying samples from all 103 languages, the precision of the baseline is always slightly higher than that of the enriched model, but this comes at the cost of a significantly lower recall and F1 score.

On the first task, the 32 dimensions model performed better although the difference between all of them was very narrow. For the second task the 64 dimensions had better results that were also slightly more significant.

Source	Baseline (16)	Enriched (16)	Baseline (32)	Enriched (32)	Baseline (64)	Enriched (64)
Tatoeba	86.05%	87.98%	86.53%	88.94%	85.57%	88.94%
Transcriptions	57.33%	82.66%	50.00%	84.66%	52.66%	84.00%
Caterina Bonan	61.53%	94.87%	61.35%	94.87%	56.41%	94.87%
Simmen Kréyòl	66.66%	91.67%	58.33%	94.44%	63.88%	94.44%
Chansons	63.63%	90.90%	63.63%	90.90%	72.72%	90.90%
Wikipedia	83.33%	83.33%	83.33%	83.33%	83.33%	83.33%
All datasets	71.37%	87.73%	68.87%	88.95%	68.91%	88.75%

Table 3: Accuracy of each model when classifying Guadeloupean Creole sentences from each of the sources

	Baseline (16)	Enriched (16)	Baseline (32)	Enriched (32)	Baseline (64)	Enriched (64)
Precision	96.14%	94.07%	96.27%	91.04%	96.28%	95.59%
Recall	71,37%	87.73%	68.87%	88.95%	68.91%	88.75%
F1 score	81.92%	90.79%	80.19%	89.96%	80.33%	92.04%

Table 4: Precision, recall and F1 score for Guadeloupean Creole of each model when classifying sentences from 103 languages from the enriched dataset

6 Error Analysis

We extracted the most common errors made by the system. Table 5 shows which languages were most commonly assigned to real GC samples. As expected, French was the most common misclassification in all but one models, in which it was the second most common. There is no clear second nor third place but it is worth noticing the constant appearance of Austronesian languages. On the other hand, Table 6 shows which languages were most commonly misclassified as CG. In this case there is no clear first place, although Spanish seems to be the most consistent error, however the distribution among the misclassified languages seems to be more uniform.

Model	Misclassification	Error	Misclassification	Error	Misclassification	Error
Baseline(16)	Kapampangan	16%	French	14%	Tagalog	6%
Enriched (16)	French	20%	Spanish	10%	Waray	8%
Baseline (32)	French	18%	Kapampangan	9%	Irish	5%
Enriched (32)	French	24%	Waray	11%	Esperanto	11%
Baseline (64)	French	18%	Kapampangan	14%	Interlingue	5%
Enriched (64)	French	22%	Esperanto	9%	Spanish	7%

Table 5: Three most common errors when classifying Guadeloupean Creole as another language and how much of the total number of errors each of them represents

Model	Real Language	Error	Real Language	Error	Real Language	Error
Baseline(16)	Spanish	14%	Low Saxon	7%	French	7%
Enriched (16)	Spanish	15%	French	11%	Finnish	11%
Baseline (32)	Breton	15%	Cebuano	15%	Lojban	15%
Enriched (32)	Kotava	14%	Spanish	12%	Portugues	12%
Baseline (64)	Spanish	23%	Lojban	15%	Interlingue	15%
Enriched (64)	Lojban	10%	Spanish	10%	Esperanto	10%

Table 6: Three most common errors when classifying another language as Guadeloupean Creole and how much of the total number of errors each of them represents

7 Conclusion

We produced a reliable Language Identification tool for Guadeloupean Creole and proved that enriching the training of the model with few but diverse sources of the language helps to improve the performance. We also produced a public version of our models along with a python wrapper and a terminal tool for ease of use[2]. There is still much room for improvement both in the design and processing of a richer set of GC examples as well as on the settings of the classification model. We hope that this article will motivate other researchers to work on the implementation of NLP tools of under-resourced languages.

[2]See `https://gitlab.com/williamsotomartinez/gclit/`

References

Armand, J., Grave, E., Bojanowski, P., and Tomas, M. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Bernabé, J. (2001). *La graphie créole*. Ibis rouge.

Bonan, C. (2018). Online corpus of guadeoupean creole. `https://caterinabonan.com/corpus-of-guadeloupean-creole/`Visited on 18/06/2020.

Brown, R. D. (2013). Selecting and weighting n-grams to identify 1100 languages. In *International Conference on Text, Speech and Dialogue*, pages 475–483. Springer.

Colot, S. and Ludwig, R. (2013). Guadeloupean and martinican creole. In Michaelis, S. M., Maurer, P., Haspemath, M., Huber, M., and Revis, M., editors, *The Survey of Pidgin and Creole Languages: Volume 2*. Oxford University Press.

Delumeau, F. (2006). *Une description linguistique du Creole Guadeloupéen dans la perspective de la genération automatique d'enonces.* PhD thesis, Université de Nanterre - Paris.

Glaude, H. (2013). Corpus Créoloral. oai:crdo.vjf.cnrs.fr:crdo-GCF, SFL Université Paris 8 - LLL Université Orléans.

Hazaël-Massieux, M.-C. (1993). Ecrire en créole(oralité et écriture aux antilles). *Journal of French Language Studies*.

Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Ludwig, R., Montbrand, D., Poullet, H., and Telchid, S. (1990). Abrégé de grammaire du créole guadeloupéen. *Dictionnaire créole français (Guadeloupe), avec un abrégé de grammaire créole et un lexique françaiscréole*, pages 17–38.

Millour, A. and Fort, K. (2018). Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen. In *CCURL 2018* , Miyazaki, Japan.

Nakatani, S. (2010). Language detection library for java. Software available at `http://code.google.com/p/language-detection/` (last updated on March 2014).

Sangeeth, K. (2017). Whatthelang. Software available at `https://github.com/indix/whatthelang`.

Sites, D. (2013). Compact language detector 2. Software available at `https://github.com/CLD2Owners/cld2` (last updated on August 2015).