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

**Raj Dabre**
CFILT
IIT Bombay
prajdabre
@gmail.com

**Aneerav Sukhoo**
Central Informatics Bureau
Quatre Bornes, Mauritius
aneeravsukhoo
@yahoo.com

**Pushpak Bhattacharyya**
CFILT
IIT Bombay
pushpakbh
@gmail.com

## Abstract

We present, in this paper, our experiences in developing Statistical Machine Translation (SMT) systems involving English, French and Mauritian Creole, the languages most spoken in Mauritius. We give a brief overview of the peculiarities of the language phenomena in Mauritian Creole and indicate the differences between it and English and French. We then give descriptions of the developed corpora used for the various MT systems where we also explore the possibility of using French as a bridge language when translating from English to Creole. We evaluate these systems using the standard objective evaluation measure, BLEU. We postulate and through an error analysis, indicated by examples, verify that when English to French translations are perfect, the subsequent translation of French to Creole results in better quality translations than direct English to Creole translation.

## 1 Introduction

Mauritius[1] is an island nation in the Indian Ocean about 2000 km off the south-east coast of the African continent. The population of Mauritius is approximately 1.3 million and it is the 151st most populated country in the world. While English language is the official language, French language is spoken to a greater extent and the Mauritian Creole (henceforth called creole) is the most common language used by the majority of the population. In addition, since the people of Mauritius have Indian ancestors, Indian languages such as Marathi, Hindi and Bhojpuri (a Bihari language) are also spoken amongst the populace. Mauritius is an extremely popular tourist spot and attracts many thousands of people every year and has such involves interaction with the local inhabitants. Although English is the main language, lack of knowledge of the Mauritian Creole (which is based on French) does lead to a gap in communication.

Various organisations have been actively involved in give the Mauritian Creole a new dimension, whereby a trilingual dictionary has been prepared by the University of Mauritius and a bilingual dictionary has also been provided online by Ledikasyon pu Travayer[2] on its website. There is room for further research in the area to come up with a dictionary for technical terms. Computational Linguistics can help to take the Mauritian Creole language to greater heights.

Google provides translation services between many languages but Mauritian Creole is not one of them which prompted us to begin the development of machine translation systems to achieve this end. To the best of our knowledge this is the first of its kind work involving translation to and from Mauritian Creole on such a large scale. We hope that this work will attract and help researchers in the development of machine translation systems involving a variety of Creole languages.

[1] http://en.wikipedia.org/wiki/Mauritius

[2] http://www.lalitmauritius.org/dictionary.php

### 1.1 Related Work

Sukhoo et al. (2014) had developed a basic English Creole SMT system with a very small amount of parallel corpus (10000-13000 lines including many dictionary words). They manage to get simple sentences translated with reasonable quality but fail when longer sentences (more than five words) are tested. This corpus, after considerable augmentation was used in our experiments. Significant work has been done in using bridge languages (Bertoldi et al. (2008), Utiyama et al. (2007)) to improve translation quality. In many of these works, they develop translation systems from languages A to C using B, where A-B and B-C are resource rich, by either synthesizing new phrase tables or modifying existing corpora. In our case, we had a huge English French corpus but a very small French Creole corpus. Moreover no linguistic processing modules for Creole exist, which would help in reducing data sparsity. Thus we decided to adopt the "transfer method" described by Wu et al. (2009). The remainder of the paper involves a chapter on Linguistic phenomena in Mauritian Creole; describing peculiarities and how it is different from English and French; followed by a chapter describing the corpora used, the systems developed, the results and error analyses via examples finally leading to the conclusion and future works.

### 1.2 Purpose of the work

The outcome of our work is a set of online translation systems aiming to:

1. Assist young students to learn English as a second language.
2. Help tourists to learn the basics of the Creole language so as to enhance communication with the local community. This can provide better enjoyment during their stay.
3. Assist expatriates and foreign business people to interact with the people of Mauritius.
4. Assist the local people, who do not master the English language to be able to translate and understand English texts, which are available in huge amount online as well as from other sources.

## 2 Mauritian Creole

The Mauritian Creole is spoken in Mauritius and Rodrigues islands. A variation of the language is also spoken in Seychelles. Mauritius was colonized successively by the Dutch, French and English. Even though the English took over the island from the French in the early 1800, French remained as a dominant language and as such Creole language shares many features with French.

### 2.1 Similarities with French

The same alphabets are used in both cases and they are pronounced in a similar manner. In addition, some words are written and pronounced in the same way. These include words as per the table below (the English translation is also shown in the 3rd column):

| French | Creole | English |
|---|---|---|
| avion | avion | aeroplane |
| bon | bon | good |
| gaz | gaz | gas |
| bref | bref | brief |
| pion | pion | pawn |

**Table 1: Words with similar orthography in French and Creole**

One must note that, in French there is a heavy usage of accents while writing which is absent in Creole. Many words are pronounced similarly in French and Creole, but the grapheme is different. These include:

| French (Fr) | Creole (Cr) | English (En) |
|---|---|---|
| mauvais | move | move |
| confort | konfor | comfort |
| méditation | meditasion | meditation |
| insecte | insekt | insect |
| condition | kondision | State, terms or provision |

**Table 2: Words with same pronunciation but different orthography**

### 2.2 Creole Grammar

The grammar of Creole has been published in 2011 (Police-Michel, Carpooran and Florigny, 2011). The Mauritian Creole language has sentences with structure of Subject-Verb-Object as in the case of English and French language. Some differences with English language can be noted as follows:

1. Adjectives are sometimes moved after the object:
   "The brown bird" is translated as: **Zwazo maron-la**
   Here: "maron" for "brown" is moved after the object (Zwazo). "The" is moved at the end ("la"). However, the French translation follows the same pattern as Creole, e.g. "L'oiseau maron"

2. Difference between singular and plural: "There are many birds" is translated as:
   **Ena boukou zwazo laba**
   Here: The plural form does not take "s".
   The word "boukou" indicates "many" and therefore, it can be deduced that there are many birds. In French, the translated sentence is "Il y a beaucoup d'oiseaux là-bas"
3. Dropping of verb:
   "He is bad" is translated as:
   **Li move**
   Here: "He" is translated to "Li" and "bad" to "méchant". The verb "is" is dropped. In French, the translated sentence becomes "Il est méchant", where the verb is retained.

This concludes the explanation of Creole linguistics. Due to the similar SVO syntax and limited morphological complexity of English, French and Creole, translation between them becomes tractable even which working without linguistic processing and small corpora. We now describe our experiments and systems developed.

## 3 Experiments

We begin by listing the various SMT systems developed. We have used only phrase-based methods due to the lack of linguistic processing modules for Creole.

### 3.1 Systems developed

All the systems developed and hosted are given below:

1. Direct English to Creole
2. Direct Creole to English
3. Direct French to Creole
4. Direct Creole to French
5. English to Creole using French as bridge

The front-end of our system was developed using a combination of HTML and DJANGO (Python). The SMT models, at the back end, are provided as services accessible by XMLRPC.

### 3.2 Corpora details

Plenty of parallel corpus is available for English-French however we had to manually, from scratch, create parallel corpus for Creole-French and Creole-English. These sets of corpora were developed by the sole Creole speaker in our group over a period of six months. A significant part of the corpora consists of dictionary words which ensure basic word substitution. Table 3 gives the details of the corpora used.

| Language pair | #lines | #words (L1-L2) | Source |
|---|---|---|---|
| En-Fr | 2000000 | 127405-147812 | Europarl |
| En-Cr | 25010 | 16294-17389 | Manually created |
| Fr-Cr | 18354 | 13769-13725 | Manually created |

**Table 3: Corpora Details**

In the French-Creole corpus 11,208 entries were just dictionary words and thus the number of parallel sentences was just around 7146. Similarly for the English-Creole corpus 11,423 entries were dictionary words making the number of parallel sentences around 13,795. For creole we had a monolingual corpus of around 100k lines.

### 3.3 Training and development

For purposes of training we used IBM models (Brown et al., 1993) implemented in GIZA++ for alignment and Moses (Koehn et al., 2007, 2003) for phrase extraction and decoding. Standard parameters were used. Training procedure following which the phrase tables and reordering tables were binarized in order to ensure on demand memory loading thereby reducing the requirement of RAM. The services for each language pair are hosted using mosesserver. The English-French system took the maximum training time; a total of 24 hours, including the time for tuning. All the other systems took around 5 minutes of training. Due to lack of corpora we did not perform tuning for these systems. For each system the target side corpus was used for generating the language model. For creole we used the monolingual corpus for language modelling. Tuning was done using MERT.

### 3.4 Using French as a bridge language

We used the "transfer method" or "sentence translation strategy", proposed by Utiyama and Isahara (2007) and described by Wu and Wang (2009), to translate from English to Creole using French; wherein the first translated from English to French using the En-Fr system and then from French to Creole using the Fr-Cr system. This method is only applicable when either both the En-Fr and Fr-Cr systems or only the En-Fr system is of high quality. The main idea is that French is close to Creole and thus an SMT system built on a small corpus would suffice for good translations. As long as the English-French

SMT system gives good translations the resulting Creole translations can be expected to be good.
The Moses decoder allows for n-best translations to be generated as outputs which we exploit to obtain improvements in translations. When Moses translates a sentence it uses 8 main features for decoding. These are: n-gram language model probability of the target language (1 feature), two phrase translation probabilities (both directions, 2 features), two lexical translation probabilities (both directions, 2 features), a word penalty (1 feature), a phrase penalty (1 feature), a linear reordering penalty (1 feature). Each of above is a feature for decoding procedure (Hoang et al., 2007; Koehn et al., 2003). The decoder gives values for each feature for a translation candidate. The final score of the translation is a weighted sum of the values of the above mentioned features. The weights are obtained using MERT.

Let "**f**" denote the source language, "**p**" the pivot language and "**e**" the target language. Let $L_1$-P denote the system that translates from source to pivot and P-$L_2$ the system that translates from pivot to target. In our case $L_1$ is English, P is French and $L_2$ is Creole. The sentence translation strategy is:

1. Translate source language sentence "f" into "N" intermediate sentences using $L_1$-P system.
   a. f → $p_1$, $p_2$, $p_3$ .... $p_N$
   b. 8 values of feature functions per $p_i$: $h^p_{i1}$, $h^p_{i2}$,....., $h^p_{i8}$
2. Translate each candidate into "N" target language sentences (e) using P-$L_2$ system
   a. $p_i$ → $e_{i1}$, $e_{i2}$, $e_{i3}$, ... $e_{iN}$
   b. 8 values of feature functions per $e_{ij}$: $h^e_{ij1}$, $h^e_{ij2}$,....., $h^e_{ij8}$
3. Score ***N*N*** target translations using feature values and Moses tuning parameters
   a. Each feature has a weight
   b. Weights obtained by tuning via MERT
   c. $\boldsymbol{\lambda^p_m}$ (m=1 to 8): Parameter of $L_1$-P system
   d. $\boldsymbol{\lambda^e_m}$ (m=1 to 8): Parameter of P-$L_2$ system
4. Select the sentence $e_{ij}$ with the highest score.

The formula for scoring is:

$$S(e_{ij}) = \sum_{m=1}^{8} (\lambda^p_m * h^p_{im} + \lambda^e_m * h^e_{ijm})$$

We experimented with 2 values of "N" where N=1 and N=10. The above method has a time complexity of ***O(N*N)*** and hence we refrained from going for any higher values of "N". Using this mechanism we tested sentences to validate the following hypothesis: "For simple sentences the translation of English to Creole using French as a bridge is much better than direct English to Creole translation". Although the Fr-Cr system is built from a relatively small corpus, we believe in the validity of this hypothesis because French is much closer to Creole than English.

### 3.5 Testing and Results

For the purpose of testing an additional 142 sentence triplets were created, one each for English, French and Creole. These are sentences with more than 10 words per sentence on an average. The number of OOV's was more in these sentences. Extra 142 simple (short) sentence pairs for English and Creole were created to verify our hypothesis mentioned above. These contained an average of 5 words per sentence with relatively lesser number of OOV's compared to the earlier sentences. We evaluated the quality using BLEU (Papineni et al. 2002). The scores are given in Table 4 below. In the table "easy" indicates that the simple (short) sentence pairs were used for testing whereas "hard" indicates that the longer ones were used for testing. For the Creole to English, Creole to French and French to Creole systems the pivot language mechanism was not applicable since effective pivot languages were not available.

| **Language Pair** | **BLEU** |
|---|---|
| En-Cr (direct, hard) | 12.90 |
| En-Cr (direct, easy) | 25.31 |
| **En-Cr (bridge (N=1), hard)** | **9.44** |
| **En-Cr (bridge (N=10), hard)** | **10.96 (increase compared to N=1)** |
| **En-Cr (bridge (N=1), easy)** | **26.12 (increase)** |
| **En-Cr (bridge (N=10), easy)** | **29.77 (increase)** |
| Cr-En (direct, hard) | 17.58 |
| Cr-En (direct, easy) | 22.58 |
| Fr-Cr (direct, hard) | 15.97 |
| Cr-Fr (direct, hard) | 14.54 |

**Table 4: BLEU scores of systems**

As expected, the English to creole translation using French as bridge yields a BLEU score higher than the direct translation for simple (short) sentences. This improvement is higher when the number of intermediate translations is increased from N=1 to N=10 (from 25.31 (direct) to 26.12 to 29.77). This is because when N=10

we may have a better intermediate French translation as compared to when N=1. The limitation is that when N=10 the time taken is 100 times more than when N=1. The increase is not observed for "hard" sentences. This is because their translations from English to French (as intermediates) were not of good quality and this led to a multiplicative degradation when translating the intermediate French sentence to Creole. However there is an improvement in BLEU from 9.44 to 10.96 when N is increased from 1 to 10.

We did significance testing using bootstrapping sampling and also performed subjective analysis for the bridge translations to verify that the increase in BLEU was not coincidental. We also observed the effects of the change in BLEU with increase in corpus size used for training for En-Cr and Cr-En which is given in the table 5 below. The BLEU scores are different here because the test set used was different from the 142 sentences mentioned before. The main difference between the corpus at 20000 size and 25000 is the increase in the number of non-dictionary sentences which indicates that even with small but good corpus size increments we can achieve tremendous improvement in quality.

| No of training sentences | En-Cr (direct) | Cr-En |
|---|---|---|
| 5,000 | 5.36 | 7.47 |
| 10,000 | 6.17 | 8.57 |
| 15,000 | 6.07 | 9.16 |
| 20,000 | 6.25 | 9.69 |
| 25,000 (full) | **7.71** | **11.64** |

**Table 5: Change in BLEU with corpus size**

### 3.6 Example Translations and Error Analysis

For each language pair we give below examples of good and bad translations.

#### 3.6.1 En-Cr

1. **Input:** They all had big ears and long legs.
   **Direct Creole:** zot tu ti ena gran zorey e longay mole.
   **French (Bridge):** ils avaient tous grandes oreilles et long jambes.
   **Resulting Creole:** zot ti ena tou gran zorey e long lazam.
   **Evaluation:** In this case, the French translation is of high quality and the Creole through bridge language is equally of high quality and is preferred as compared to the direct Creole translation.

2. **Input:** My name is John.
   **Direct Creole:** mo apel li.
   **French (Bridge):** mon nom est john.
   **Resulting Creole:** mo apel zan.
   **Evaluation:** Because the French translation is correct the Creole translation as compared to the direct Creole translation, whereby the name "John" has not been translated.

3. **Input:** There are birds like no other birds because they are not real.
   **Direct Creole:** ena bann zwazo kuma okenn lezot zwazo parski zot pa vre.
   **French (Bridge):** il y a des oiseaux comme aucun autre oiseaux car ils ne sont pas vraies.
   **Resulting Creole:** ena bann zwazo kuma okenn lot zwazo parski zot pa vraies.
   **Evaluation:** The direct Creole translation is correct. The French translation is correct, but the only issue in the Creole through bridge translation is the word "vraies" in French, that was left as it is given that it is not a stem word. The small size of the French-Creole parallel corpora has been a limitation to this translation. The translation of the other words, including the order, is also correct.

#### 3.6.2 Cr-En

1. **Input:** mo pa kontan fason to koz ar mwa
   **English:** i don't like the way you talk with me
   **Evaluation:** Correct translation.

2. **Input:** boukou touris pou vizit Moris sa lane la
   **English:** many tourist to visit to mauritius this year
   **Evaluation:** The correct tense of the verb "to visit" is "will visit", otherwise the rest of the sentence is correct.

3. **Input:** alor li ti dir sa grenouy la , "si to pou donn mwa enn zafer , mo pou fer tou seki to dimande"
   **English:** so he said to the frog, if you will give me a thing, i will do what you ask
   **Evaluation:** Almost correct translation, apart from the elimination of the quotes. "a thing" should have been "something".

#### 3.6.3 Fr-Cr

1. **French:** je voudrais regarder un film (I want to watch a movie)
   **Creole:** mo anvi get enn fim

**Evaluation:** This translation is perfect.

2. **French:** il y a des tigres et des lions dans la forêt (There are tigers and lions in the forest)
   **Creole:** ena bann tigres ek bann lions dan afore
   **Evaluation:** Here the Creole sentence sounds like "There are the tigers and the lions in the forest". The word "bann" (the) should be absent.

3. **French:** pourquoi êtes-vous en colère contre moi? (Why are you angry with me?)
   **Creole:** kifer eski zot ankoler kont mwa?
   **Evaluation:** This translation is particularly bad. The correct is: "kifer to ankoler ar mwa". The system was unable to correctly translate "êtes-vous" (are you). This we believe is a major limitation of small corpus size.

#### 3.6.4 Cr-Fr

1. **Creole:** mo anvi manz enn dipin (I want to eat bread)
   **French:** je veux manger un pain
   **Evaluation:** This translation is perfect.

2. **Creole:** kifer to ankoler ar mwa (Why are you angry with me?)
   **French:** pourquoi vous en colère avec moi.
   **Evaluation:** The translation is almost perfect except that it sounds like "Why you angry with me?". "are" is missing as "vous" should have had "êtes" with it.

3. **Creole:** sa lane la Moris finn gagn boukou touris (This year Mauritius had a lot of tourists.)
   **French:** cette année de l'île maurice a eu beaucoup de tourists
   **Evaluation:** Here the French translation sounds like: "this year's maurice island had a lot of tourists". Here "la Moris" is a named Entity which actually means Mauritius and its translation as "l'île maurice" is acceptable. The "de" after "année" is incorrect.

### 3.7 Discussion

The examples given above should indicate that survival sentences are translated with a very high quality in most cases. Common mistakes are those involving incorrect tenses, dropping of words although they are present in the corpus, adding articles like "bann" (the) for each noun and the inability to handle named entities. The main reason for this we narrow down to the following:

1. Poor language model for creole due to relatively small corpus. The creole monolingual corpus was around 100k lines but not clean since it was collected from a variety of sources. Also it is common knowledge that a corpus of atleast 1 million lines is good for language modelling.
2. Insufficient decoding options due to lack of evidence in corpora (small corpus size) and hence phrase table.
3. Lack of factors (linguistic cues) leading to surface forms being un-translated.
4. Lack of lower/upper case information which can help recognize named entities; which is again due to small corpus size. (We lowercase then translating from/to Creole as truecasing is ineffective due to a small corpus.)
5. If a pivot language is used then the final quality depends on the quality of the intermediate translations which, if bad, lead to poorer target translations.

This gives sufficient reason for us to look deeper into the decoding procedure and perhaps fine tune it for our purposes. Also the development of a larger corpus is a necessary activity. Finally more linguistic phenomena have to be studied and uncovered which would eventually lead to analysis modules for Creole.

## 4 Conclusion and Future Work

We have presented our experiences in developing statistical machine translation systems for Mauritian languages. As can be seen, even with small amount of corpus, we are able to get reasonable quality translations which we believe will gradually improve as the dictionary and corpora size increases. We also validated our hypotheses that French, being closer to Mauritian Creole, is a good bridge language and translation from English to Creole via French, rather than directly, gives better translations when the English to French translation is near about perfect. In the future we plan to experiment with factored models (Koehn et al., 2007), which we were not able to use due to the lack of linguistic processing modules for Creole. We also plan to experiment on translating Creole to English using French as a bridge language when our Creole French corpora size becomes large. Naturally this would lead to work on the lesser spoken languages in Mauritius as also on other Creoles.

## Reference


Aneerav Sukhoo, 2014. *Translation between English and Mauritian Creole: A Statistical Machine Translation Approach.* IST-2104, Mauritius, May, 2014.

Peter E Brown, Stephen A. Della Pietra. Vincent J. Della Pietra, and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimationn*. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst. 2007. *Moses: Open Source Toolkit for Statistical Machine Translation*, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. *Statistical phrase-based translation*. Association for Computational Linguistics 2003.

Philipp Koehn and Hieu Hoang. 2007. *Factored translation models*. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 868-876, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translation*, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002, pp. 311-318.

Nicola Bertoldi, Madalina Barbaiani, Marcello Federico, and Roldano Cattoni. 2008. Phrase-based statistical machine translation with pivot languages. Proceeding of IWSLT, pages 143–149.

Masao Utiyama and Hitoshi Isahara. 2007. *A comparison of pivot methods for phrase-based statistical machine translation.* In Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 484–491, Rochester, New York, April. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2009. *Revisiting pivot language approach for machine translation*. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 154–162, Suntec, Singapore, August. Association for Computational Linguistics.