Parsing for Mauritian Creole using Universal Dependencies

Neha Ramsurrun, Rolando Coto-Solano, Michael Gonzalez
Dartmouth College
{neha.ramsurrun.23, rolando.a.coto.solano, michael.j.gonzalez.23}@dartmouth.edu

Abstract

This paper presents a first attempt to apply Universal Dependencies (De Marneffe et al., 2021) to train a parser for Mauritian Creole (MC), a French-based Creole language spoken on the island of Mauritius. This paper demonstrates the construction of a 161-sentence (1007-token) treebank for MC and evaluates the performance of a part-of-speech tagger and Universal Dependencies parser trained on this data. The sentences were collected from publicly available grammar books (Syea, 2013) and online resources (Baker and Kriegel, 2013), as well as from government-produced school textbooks (Antonio-Françoise et al., 2021; Natchoo et al., 2017). The parser, trained with UDPipe 2 (Straka, 2018), reached F1 scores of UPOS=86.2, UAS=80.8 and LAS=69.8. This fares favorably when compared to models of similar size for other under-resourced Indigenous and Creole languages. We then address some of the challenges faced when applying UD to Creole languages in general and to Mauritian Creole in particular. The main challenge was the handling of spelling variation in the input. Other issues include the tagging of modal verbs, middle voice sentences, and parts of the tense-aspect-mood system (such as the particle *fek*).

Keywords: Treebank, Mauritian Creole, Universal Dependencies

1. Introduction

This paper presents a first attempt to create a Universal Dependencies (De Marneffe et al., 2021) treebank for Mauritian Creole (glottolog `mori1278`, henceforth MC), a French-based Creole spoken on the island of Mauritius, located off the eastern coast of Madagascar. This work also includes the training of a parsing model to automatically identify the parts of speech (POS) and universal dependencies for sentences in MC. While there have been an increasing number of Universal Dependencies treebanks for other low-resource languages, such as the Indigenous languages of the Americas (Ferraz Gerardi et al., 2021; Rueter et al., 2021; Vasquez et al., 2018; Tyers and Henderson, 2021; Park et al., 2021; Coto-Solano et al., 2021; Wagner et al., 2016; Thomas, 2019) and languages from Subsaharan Africa (Dione, 2021; Kahane et al., 2022; Aplonova, 2018; Aplonova and Tyers, 2017; Kahane et al., 2023) and Polynesia (Karnes et al., 2023), there is not much previous work on Creole languages. In fact, the English-based Creole Naija from Nigeria is the only Creole language that has a Universal Dependencies treebank so far (Caron et al., 2020), although there is ongoing work for other French-based creoles like Martinican Creole (Mompelat et al., 2022) and Guadeloupean Creole (Millour and Fort, 2018).

MC is a language that has only recently developed a standardized spelling for its written form and has being taught in primary schools since 2012 as an optional language (Harmon, 2015). It originated in the days of slavery and the French colonial era in the 18th century and is currently spoken by around 1.3 million people both in Mauritius and in diaspora communities in Europe and Australia (Baker and Kriegel, 2013). Some features of the language include its SVO word order, preverbal tense, mood and aspect (TAM) markings, and lack of passive constructions (Syea, 2013, 13).

Since MC has just begun being taught in schools in Mauritius, there is no commonly adhered to system of spelling. There is a standard of spelling supported by the government, the *Lortograf Kreol Morisien* created by the Akademi Kreol Morisien (Carpooran et al., 2011), which is also used in dictionaries for the language (Carpooran, 2011) and in its official grammar (Police-Michel et al., 2012). However, this orthography is not in widespread use amongst speakers (Saarinen, 2016; Millour and Fort, 2020). Moreover, children in primary education can choose not to take classes in MC and instead study Asian languages that are heritage languages to the Mauritian population, such as Hindi or Tamil (Harmon, 2015; Auckle, 2023). This makes it so that many students have very limited contact with the official orthography of the language. The spelling used in this work is based on Anand Syea's book *The Syntax of Mauritian Creole* (2013), which roughly follows the guidelines in *Lortograf* (see section 2.1 for some differences between the two).

There has been previous research on NLP for MC, particularly for machine translation between MC and English or French (Dabre and Sukhoo, 2022; Pudaruth et al., 2021; Boodeea and Pudaruth, 2020; Dabre et al., 2014; Sukhoo et al., 2014; Pudaruth et al., 2013). There has also been work on speech recognition (Macaire et al., 2022; Noormamode et al., 2019), tokenization (Petrov et al., 2023), language identification (Adebara et al.,

20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

2023, 2022), stemming (Gobin-Rahimbux et al., 2023), and the crowdsourcing and compilation of corpora (Bastien et al., 2022; Millour and Fort, 2020). Interestingly, there is also NLP work on language generation, specifically the generation of Mauritian *Sega* lyrics (Bhaukaurally et al., 2012). This amount of research is unusually high amongst Creoles, which have historically not been a focus of NLP research, and which to this day are heavily under-resourced (**?**).

In this paper, we seek to share how we built a Universal Dependencies treebank for MC and how we trained a parser using UDPipe 2 (Straka, 2018) to identify POS and universal dependencies in MC. We hope that our work will help create more NLP resources for MC, and also help with language instruction.

2. Methodology

In subsection 2.1, we discuss the sources of the unlabelled data that was included in the corpus. In subsection 2.2, we discuss our process for manually tagging the unlabeled sentences, utilizing the CoNLL-U Format, Universal POS Tags and Universal Dependency Relations. Finally, in subsection 2.3, we discuss the algorithms used for the training of our parser, as well as the metrics used to analyze its performance.

2.1. Unlabeled Data Collection

We gathered sentences from multiple sources. Our main resource was the textbook *The Syntax of Mauritian Creole* (Syea, 2013), which contributed the bulk of the sentences that were ultimately tagged. These sentences covered a wide range of syntactic constructions, from simple equative sentences to sentences with relative clauses. In addition to the textbook, we also used sentences from online resources like the MC entry in the APiCS Atlas (Baker and Kriegel, 2013), as well as printed schoolbooks for Mauritian students (Antonio-Françoise et al., 2021; Natchoo et al., 2017). In total we collected 161 sentences. They contained 1007 tokens, with a total of 286 unique tokens, and had an average length of 6.0 ±2.3 tokens.

We tokenized the sentences manually. This was not problematic given the highly isolating nature of MC morphology. For example, tokens like *tifi-la* “the girl" were split into two tokens, following Syea (2013, 223) and contra Police-Michel et al. (2012, 6). A few multimorphemic tokens were not split. For example pronouns like *bann-la* “they", which could have potentially been split into the plural word and the third person pronoun, were not split so that they could be semantically different from the singular third person pronoun *la*. In this we deviate from the usage in Syea (2013, 86) (e.g. *Bann la pe bwar* “They are drinking") and follow the orthography in Police-Michel et al. (2012, 63) (e.g. *bann-la mem gete* “they themselves see").

There are a few additional points where the two orthographies differ. For example, the official orthography is mesolectal, in that it presents a graphic form that is neither completely French, nor completely creole[1] (Auckle, 2023; Carpooran, 2011). For example, when it represents the phoneme /u/, it keeps a spelling that is etymologically related to its French equivalents (e.g. *koumans* “to begin" (Police-Michel et al., 2012, 22)). On the other hand, Syea (2013) uses a more phonetic spelling for such words (e.g. *kumans* “to begin" (ibid, 60)). We used the latter spelling in our tagged sentences.

2.2. POS Tagging and Universal Dependencies

Once the sentences were collected, they were tagged manually using the CoNLL-U format, Universal Dependencies relations and Universal Parts of Speech. The tagging was done using UD Annotatrix (Tyers et al., 2018). Figure 1 shows an example of a sentence tagged in Mauritian Creole, following the format that was used to train the parser. Section 5 explains some of the challenges found when applying the UD rules to the MC data.

```
#source = The Syntax of Mauritian Creole
#id = syntaxmc-pg26-ex5a
#text = Mari pa kwi pwason .
#text_en = Mary does not cook fish.
1       Mari    _       PROPN   PROPN   _       3       nsubj   _       _
2       pa      _       ADV     ADV     _       3       advmod  _       _
3       kwi     _       VERB    VERB    _       0       root    _       _
4       pwason  _       NOUN    NOUN    _       3       obj     _       _
5       .       _       PUNCT   PUNCT   _       3       punct   _       _
```

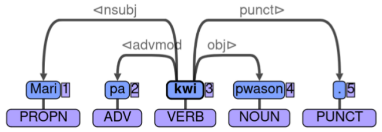


Figure 1: Example of MC sentence in CoNLL-U format, and its visualization in UD Annotatrix (Tyers et al., 2018)

[1] A *mesolectal* representation of a language variety is one that is “halfway" between the lexifier language (e.g. French) and a completely creole form. Compare this to an *acrolectal* representation, one that is more similar to French, or a *basilectal* representation, one that is the furthest away from French. An example of this in Jamaican Creole English would be the word *north*, which can be written in three ways: acrolectal *north*, mesolectal *naht*, and basilectal *naat* (Handke, 2012). In Mauritian Creole, the word “to begin" could be written as acrolectal *commence*, mesolectal *koumans* or basilectal *kumans*.

2.3. Parser Training and Evaluation

We used UDPipe 2 to train an RNN-based model for automatic parsing and POS tagging. We trained six models, with the sentences randomly split into 80% training, 10% validation and 10% test sets. The hyperparameters for training can be found in Appendix 1. The performance of the models was evaluated using three metrics: (i) F1 for the POS tags, (ii) F1 for unlabeled attachment score (UAS), and (iii) F1 for labeled attachment score (LAS).

3. Results

Table 1 shows the frequency of the UPOS tags in the dataset. The four most common parts of speech were VERB, PUNCT, PRON and NOUN; together they account for 65% of the words in the corpus. There were a total of 15 tags used; the category of “others" includes the tags NUM (n=9), CCONJ (n=2) and INTJ (n=1).

VERB	184 (18%)	ADV	45 (4%)
PUNCT	167 (17%)	PROPN	44 (4%)
PRON	154 (15%)	ADP	39 (4%)
NOUN	153 (15%)	ADJ	27 (3%)
AUX	102 (10%)	SCONJ	18 (2%)
DET	62 (6%)	Others	12 (1%)

Table 1: Frequency of UPOS tags in MC sentences.

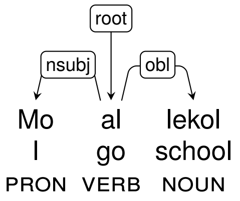
Table 2 shows the frequency of the relations found in the dataset. The relations *punct*, *nsubj*, *root* and *aux* are the most common relations. There are some relations, such as *ccomp* and *appos*, which appear relatively infrequently, but we expect them to occur more frequently as the corpus is expanded. There were a total of 29 relations used; the category of “others" includes relations like amod (n=14), iobj and nmod (n=5 each), ccomp, conj, acl:relcl, obl:tmod and nummod (n=4 each), compound:redup (n=3), appos and cc (n=2 each), and discourse, cop, compound and acl (n=1 each).

punct	167 (17%)	advmod	45 (4%)
nsubj	158 (16%)	case	39 (4%)
root	153 (15%)	nmod:poss	30 (3%)
aux	104 (10%)	xcomp	23 (2%)
obj	87 (9%)	mark	19 (2%)
det	62 (6%)	advcl	17 (2%)
obl	48 (5%)	Others	55 (5%)

Table 2: Frequency of relations in MC sentences.

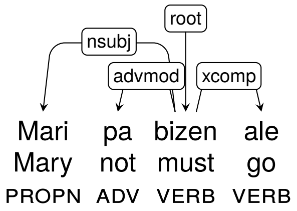
The examples below show dependency parses for three sentences of differing lengths and complexities: (a) *Mo al lekol* “I go to school", (b) *Mari pa bizen ale* “Mary does not have to go" and (c) *Okenn so fami pa finn get li* “No one from his family has seen him."

(a)



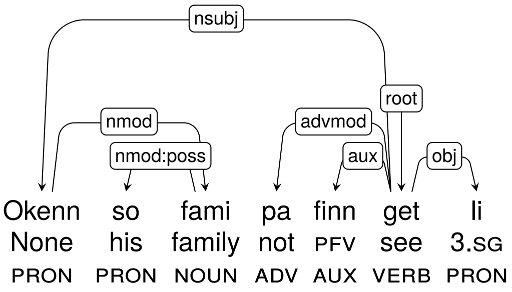
“I go to school"

(b)



“Mary does not have to go"

(c)


nsubj
root
nmod
nmod:poss
advmod
aux
obj
Okenn so fami pa finn get li
None his family not PFV see 3.SG
PRON PRON NOUN ADV AUX VERB PRON

“No one from his family has seen him"

After the data was tagged, we trained a parser using UDPipe 2, as described in section 2.3. A summary of the results for UPOS, UAS and LAS metrics are shown in figure 2. Out of the six models trained, the average UPOS was 86.2 ±4.4, the average UAS was 80.8 ±7.2, and the average LAS was 69.8 ±7.3.

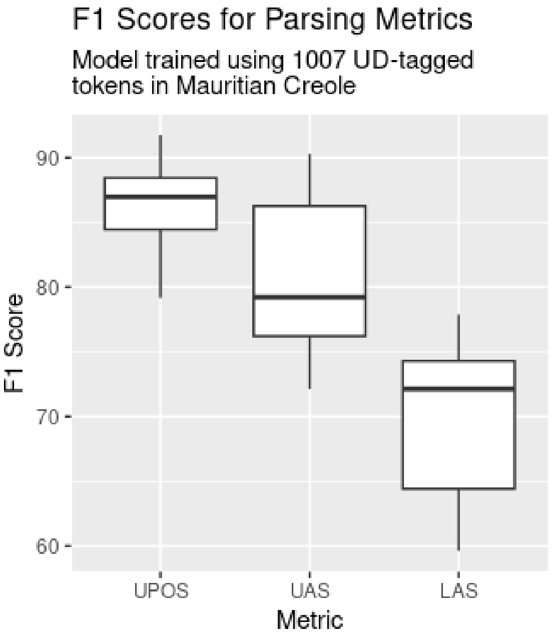


Figure 2: Results of UDPipe 2 training

In order to understand how these results compare to those of other treebanks, we calculated the average F1 for UPOS, UAS and LAS of treebanks in 75 languages, trained using the same parser (UDPipe 2) and reported by Kondratyuk and Straka (2019). The results can be see in table 3. The MC results are lower that the average for all the trees. For example, the LAS is 10 points lower (69.8 compared to 79.7 for all trees). However, the average number of sentences used to train these trees is 7544, much higher than those in the MC treebank here. In order to establish a fairer comparison, we also include data from a treebank in Lithuanian, a low-resource language at the time of the test. This treebank was trained using 154 sentences, and it is the one closest in size to our MC dataset. In this case, the MC model outperformed the Lithuanian model by an ample margin. For example, the UAS and LAS of the MC were 28 points higher. The advantage for the UPOS was not as pronounced, but MC still outperformed the Lithuanian model by 4 points. In general, the MC model appears to have an acceptable performance.

	MC	All	`lit`
Sentences	161	7544 ± 9985	154
UPOS	86.2 ± 4.4	93.7 ± 10.1	81.7
UAS	80.8 ± 7.2	84.3 ± 12.2	52.0
LAS	69.8 ± 7.3	79.7 ± 15.0	42.2

Table 3: Comparison of Mauritian Creole (MC) UDPipe 2 F1 results with those of 75 languages and Lithuanian (`lit`) (Kondratyuk and Straka, 2019), the language most similar in size to the MC corpus presented here.

There are relatively few published performance results for Indigenous languages. There are no pretrained UDPipe 2 models available for the Indigenous languages in Universal Dependencies at the time of writing (v2.12), but there are a few articles where results are available. Table 4 summarizes these results. The languages that are most directly comparable in terms of treebank size are Cook Islands Māori (155 sentences) and Yoruba (140 sentences). The MC results are much higher than those of Yoruba (e.g. UPOS$_{MC}$=86, UPOS$_{Yoruba}$=59), but lower than those of Cook Islands Māori (UPOS$_{CIM}$=93). The results for Yoruba and Cook Islands Māori, combined with the results from Lithuanian from table 3, show the high level of variability in performance that can be observed in such small datasets. The MC treebank presented here appears performs better than two of them (Lithuanian and Yoruba) but worse than Cook Islands Māori.

Another way to compare the MC results would be to find datasets that achieve similar performance to the MC model. Table 4 shows results for Western Sierra Puebla Nahuatl (Pugh et al., 2022), which has roughly similar F1 scores, but needs many more sentences to get there (939 for Nahuatl versus 161 for MC). Yupik (Park et al., 2021) and K'iche (Tyers and Henderson, 2021) also outperform the MC model, but they also need more sentences to reach that higher performance.

	Sentences	UPOS	UAS	LAS
CIM	155	93	92	86
K'iche	1433	97	91	87
Konibo	407	(NA)	84	78
Nahuatl	939	89	77	68
Yoruba	140	59	45	29
Yupik	309	93	89	82
Tr+Mund	158	42	(NA)	NA
Wo+Apur	148	58	(NA)	NA
MtC	240	(NA)	72	63
Fr+MtC	240	(NA)	81	72
MC	161	86	81	70

Table 4: Comparisons of F1 for Mauritian Creole (MC). The first group has the under-resourced languages Cook Islands Māori (CIM) (Karnes, 2023), Maya K'iche (Tyers and Henderson, 2021), Shipibo Konibo (Vasquez et al., 2018), Western Sierra Puebla Nahuatl (Pugh et al., 2022), Yoruba (Dione, 2021) and St. Lawrence Island Yupik (Park et al., 2021). The second group has results for transfer from Turkish to Mundurukú and from Wolof to Apurinã (de Vries et al., 2022). The third group has the Creole language Martinique Creole (MtC) and a Martinique Creole model with transfer from French (Mompelat et al., 2022).

The second part of table 4 has POS results for a transfer learning experiment (de Vries et al., 2022). The paper includes two Indigenous languages from Brazil which have treebanks of comparable size to our MC dataset: Mundurukú (158 sentences) and Apurinã (148 sentences). The best results for these languages were for transfer from the Turkish and Wolof models respectively. Both of them (UPOS$_{Mund}$=42, UPOS$_{Apur}$=58) were lower than the UPOS=86 for MC.

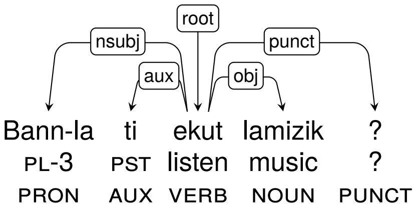
There are even fewer published performance results for the parsing of Creole languages. The only comparable one is an experiment for Martinique Creole (Mompelat et al., 2022). This experiment used 240 sentences, about double those of the MC corpus used here, but got lower F1 results (e.g. LAS$_{MtC}$=63 vrs. LAS$_{MC}$=70). The results for the Martinique Creole matched those for MC only when the authors used transfer learning from a French model. This means that the MC model presented here performs better than those for other creoles.

In summary, the model appears to have a favorable performance compared to other models of equal size for under-resourced languages.

4. Discussion

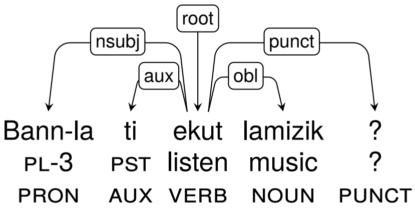
The performance of the parsing training allows us to make some generalizations about the models. For example, at an average of F1$_{POS}$=86, the model seems to be capturing parts of speech with acceptable accuracy compared to languages with a similar amount of data. Let's look at the sentence *Bann-la ti ekut lamizik?* "Did they listen to music?". Example (d.i) shows the gold-standard parse, and example (d.ii) shows an automatic parsing. Here the UPOS=100; all of the parts of speech are tagged correctly. One of the arrows, the one that connects the verb *ekut* 'listen' to the direct object *lamizik* 'music' is incorrectly marked as an oblique object, giving us a LAS=80.

(d.i)



"Did they listen to music?"

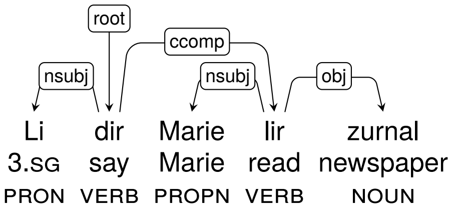
(d.ii)



(Automatic parsing: UPOS=100, LAS=80)

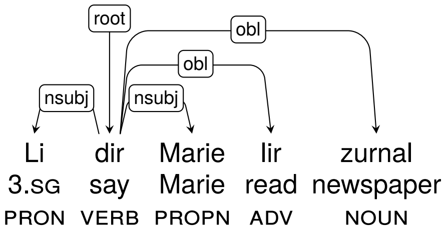
A second generalization is that the LAS is still relatively low, particularly for complex sentences. Examples (e.i) and (e.ii) show the gold-standard and an automatic parse for the sentence *Li dir Marie lir zurnal* 'He says that Marie reads the newspaper'. This sentence includes a subordinate clause, which would call for the relation ccomp. This relation only appears four times in the corpus. As expected, this sentence is not parsed correctly. The model misunderstood the subordinated verb *lir* 'reads' as an adverb and an oblique argument, which causes the other arguments to be mislabelled as well. For this example, UPOS=80 and LAS=20, a relatively low performer compared to the rest of the corpus.

(e.i)



"He says that Marie reads the newspaper"

(e.ii)



(Automatic parsing; UPOS=80, LAS=20)

It is important to note that the average results might be higher than expected because of the small size of the dataset. It might be the case that, given the low number of unique tokens, the POS tags are relatively easy to learn for the model.

5. Challenging Structures

There were several structures where there wasn't a straightforward match between the UD specifications and the MC syntax. In this section we explain some of our decisions in tagging the corpus.

Spelling. The MC official orthography is of recent creation, and it is still not used by the majority of the population. In our dataset, we tended to follow the spellings from Syea (2013), but in practice, many users of MC would not follow standardized spelling. Table 5 has examples of non-standard writing from news forums in social media. The overarching trend is the usage of French-inspired acrolectal spellings (e.g. *la guerre* versus *lager* for "war"). This is common in Creole languages, where there is a tendency to keep the orthography of words borrowed from the European language they are related to (their *lexifier* language). This comes from a dynamic tension between wanting to assert the Creole's independence (which would lead to orthographic forms divergent from the European norm), and wanting to borrow from the European language's higher status in society, as a way to enhance the Creole's own status[2] (Auckle, 2023; Fishman, 2011). While the standard spelling aspires to maximize its graphemic distance to French, many people who write MC keep the original French spelling of the borrowings in an effort to make the language look more readable and potentially more prestigious (Auckle, 2023). There is also a tendency common to contemporary written language: There are numerous abbreviated spellings in social media and electronic communications (e.g. *p* instead of *pe* "imperfective"; *c* instead of *se* "it is").

Because of this variation, the model's performance may be impacted if the input deviates from the orthography we have used. The variation is

[2] This is a tendency common to all languages undergoing standardization. English, for example, borrowed numerous Greek and Latin terms in the 17th century as a way to enhance its status as a language of science (Auckle, 2023; Hill, 2010).

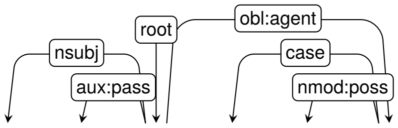
Observed form	Standardized orthography	English
Mauricien mem ki p vote ban dimun la dan sa ban post la. (Nullathemby, 2023)	*Morisien mem ki pe vot bann dimun la dan sa bann post la*.	“It's Mauritians that are voting these people into these" positions.
La guerre pou dan parlement ca cou la. (Poomen, 2023)	*Lager pu dan parlman sa kou la*.	“War will be in parliament this time".
Get nou ban weekend cuma ti été, nou ban stades trembler, et niveau ti xtra fort. (Fans, 2023)	*Get nu bann weekend kuma ti ete, nu bann stad tremble, ek nivo ti extra for*.	“Look how our weekends used to be, our stadiums tremble and the level was very high".
C enn crime contre nou patrimoine. (Fans, 2023)	*Se enn krim kont nu patrimwann*.	“It's a crime against our heritage".

Table 5: Examples of non-standard Mauritian Creole spelling in social media

not limited to social media; it is also present in formal texts. For example, the word “to get" has two common, phonetically-motivated spellings: *gayn* and *gany*, both etymologically related to the French word *gagner*. Our treebank only includes the form *gayn*, which is the closest spelling to the phonetic form of the word: [gã̃j]. However, the form *gany* is commonly observed in sources like APiCS (Baker and Kriegel, 2013), and in the Syea (2013) grammar. Moreover, there are etymological spellings derived from the French orthography, like *gagn*, which appear in official documents (Bastien et al., 2017, 2, Police-Michel et al., 2012, 13, Carpooran et al., 2011, 13). Given the limited data available for training, we have standardized all appearances of this word to *gayn*. However, we expect the parser to perform worse when a sentence would include forms like *gany* or *gagn*.

There is research on languages without standardized orthography that is directly relevant to the MC treebank. Languages like Swiss German do not have a written standard, and constant variation is the norm in their datasets. For example, Clematide et al. (2016, 63) studied a ten thousand word corpus and found the word *nächste* ‘next' written in 29 different ways, with forms such as *nächst*, *nächscht*, *nöchst*, *nögscht* and *nögst* appearing in the data. Kew (2020, 38) compiled an ASR corpus and found the words *sind sie* ‘you are' written as *sind si*, *siter*, *sind sie*, *sitr*, *send sie*, *sendsi* and *sytdihr*. There is research on how to carry out tasks like POS tagging (Hollenstein and Aepli, 2014; Scherrer et al., 2019) and G2P (grapheme to phoneme) (Schmidt et al., 2020) in such high variation environments, and future work will have to switch focus from manual standardization (such as that in section 2.1), to handling variation in the model itself and making this work more scalable.

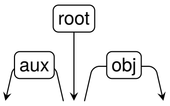
Active, Middle and Passive Voice. The sentences in our treebank are all in the active voice. MC has been described as a language that rarely uses passive constructions. These types of sentences do exist (e.g. *Li gayn koze ar so mama* “He gets shouted at by his mother" (Syea, 2013, 20); *Li finn gayn krie ar so mama* “He was shouted at by his mother"). However, these are highly infrequent. An example parse is shown in example (f).



(f)	Li	gayn	koze	ar	so	mama
	PL-3	get	shout	by	his	mom
	PRON	AUX	VERB	PREP	PRON	NOUN

“He gets shouted at by his mom"

On the other hand, the middle voice is relatively common, as in the example *Pe vann pwason* “Fish is being sold", lit “[one] sells fish"[3]. In the treebank these sentences are treated as active voice even though they have no specified subject. This is shown in example (g):



(g)	Pe	vann	pwason
	IPFV	sell	fish
	AUX	VERB	NOUN

“Fish is sold" (French: *On vend du poisson*)

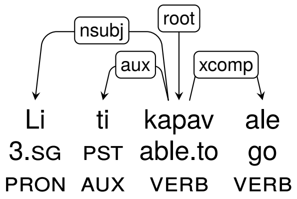
Modal Verbs. According to existing syntactic analyses (Syea, 2013), words like *bizen* “must", *kapav* “can", and *fode* “must" are modal verbs, having both verb-like and auxiliary-like properties. We can distinguish modal verbs from each other in MC

[3] This sentence is middle voice, and not passive, because it cannot take an agent as an argument. Adding an argument like *ar li* “by him" results in an ungrammatical sentence.

based on whether they are more verb-like or more auxiliary-like. We use this criterion to tag them as either VERB or AUX.

For verb-like modals, auxiliaries like *pe* “progressive marker", *ti* “past marker", *finn* “past perfective marker" can be inserted in front of them. For example, the sentence *Li ti kapav ale* “He could go" has the past tense marker in front of *kapav* “to be able to". The verb *ale* “to go" would be tagged with the relation xcomp, as shown in example (h). These verb-like modals can also exist without another main verb in the clause. For example, in the sentence *Li bizen* “He must; he has to", we tag this modal as a verb.

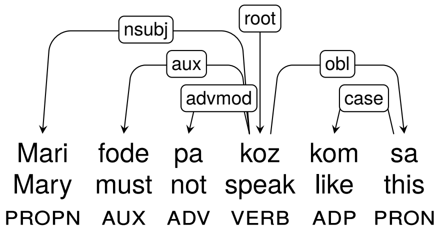
(h)



“He could go"

On the other hand, there are modals whose distribution has the same patterns as auxiliaries. One example is *fode* “must". This word is tagged as AUX because it is ungrammatical for it to be preceded by TAM markers like *ti*, and because it cannot be the only verb in the sentence. Example (i) shows a parsed example with this word.

(i)

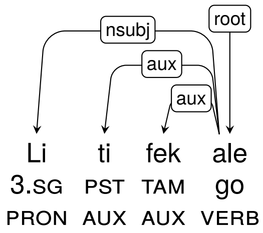


“Mary must not speak like this"

Fek. The word *fek* is usually translated as “just". Two examples of this are *Li fek ale* “He's just gone (now)", and *Li ti fek ale* “He had just gone". There are two possible parts of speech for this word: It could be an ADV modifying the verb, or it could be a TAM-marking AUX, working together with the other TAM marker *ti* (Syea, 2013, 115; Choy, 2014, 49; Police-Michel et al., 2012, 92; Adone, 1994, 43; Seuren, 1995, 534, Baker and Corne, 1986, 174).

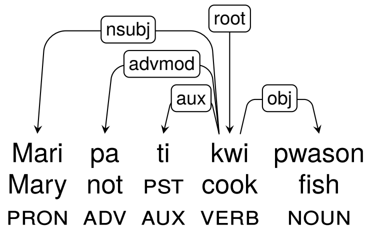
We have chosen to tag preverbal *fek* as AUX because of the tradition of understanding this as a TAM marker, and because of its distributional properties. In the example (j.i), the POS for *fek* could be either an adverb or a TAM marker. True adverbs can appear before the past marker *ti*. For example, the negative adverb *pa* “not" precedes *ti* in *Mari pa ti kwi pwason* “Mary did not cook fish"; the parse is shown in (j.ii). However, the word *fek* cannot occur in this position: **Li fek ti ale* is ungrammatical. This distribution matches that of other TAM markers, such as *finn* “past perfective" and *pe* “imperfective", which also cannot precede *ti*. Because *fek* patterns with these other TAM markers, we have chosen the AUX part of speech for it, and the aux relation for the connection between *fek* and its verb.

(j.i)



“He had just gone"

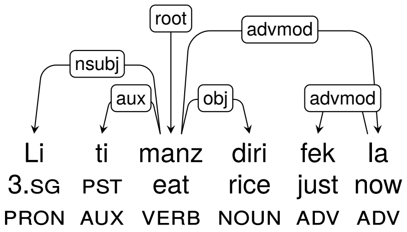
(j.ii)



“Mary has not cooked fish"

Notice that the word *fek* can have a role as an adverb in other positions. For example, in sentence (k), *Li ti manz diri fek la* “He has eaten rice just now", *fek* is modifying the word “now". Here, the relationship should be between the two adverbs.

(k)



“He has eaten rice just now"

6. Conclusions and Future Work

This paper presents a first attempt to apply Universal Dependencies to Mauritian Creole and to train a parser for Universal Dependencies for this language. The performance of the parser trained from our tagged corpus showed good levels of part-of-speech tagging (UPOS=86.2), as well as the need for improvement in the parsing of relations (UAS=80.8, LAS=69.8). Future work should include the expansion of the corpus, particularly adding longer and more complex sentences. Another future expansion of this work should consider how to compensate for alternations in spelling and other non-standard input. We hope to present this to the NLP community in Mauritius so that this can help spur more research into the structure of the language, and also contribute to generating interest amongst the general community to expand the use of Mauritian Creole into more domains of usage. We also hope that these results are useful to researchers working on NLP for other Creole languages, so that the community can begin to coalesce and collaborate on best practices for the digital description of these languages.

7. Acknowledgements

The authors wish to thank Gitanjali Seepujak for her information and clarifications about the role of Creole in the Mauritian school system.

8. Bibliographical References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for African languages. *arXiv preprint arXiv:2210.11744*.

Ife Adebara, AbdelRahim A Elmadany, and Muhammad Abdul-Mageed. 2023. Improving African Language Identification with Multi-task Learning. In *4th Workshop on African Natural Language Processing*.

Dany Adone. 1994. *The acquisition of Mauritian Creole*, volume 9. John Benjamins Publishing.

Beatrice Antonio-Françoise, Sandrine Amélia, Jersline Gaspard, Tenusha Jundoosing, Yani Maury, and Isstiac Gooljar. 2021. *Kreol Morisien - Grad 1 Volim 1*. Mauritius Institute of Education.

Ekaterina Aplonova. 2018. Development of a Bambara Treebank. *ARANEA 2018*, page 7.

Ekaterina Aplonova and Francis Tyers. 2017. Towards a dependency-annotated treebank for Bambara. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 138–145.

Tejshree Auckle. 2023. Orthography, ideology and the codification of Mauritian Creole: The implications of decreasing linguistic Abstand. *Journal of Pidgin and Creole Languages*.

Philip Baker and Chris Corne. 1986. Universals, substrata and the Indian Ocean creoles. *Substrata versus universals in creole genesis*, pages 163–183.

Philip Baker and Sibylle Kriegel. 2013. Mauritian Creole. In Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors, *The survey of pidgin and creole languages. In "The survey of pidgin and creole languages". Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press, Oxford.

Daniella Bastien, Nita Rughoonundun-Chellapermal, Ahad Mungralie, Selvavadee Pakkiri, and Jean-Noël Jolicoeur. 2017. *Ki pase la? - Grad 6 Volum 2*. Mauritius Institue of Education.

David Joshen Bastien, Vijay Prakash Chumroo, and Johan Patrice Bastien. 2022. Case Study on Data Collection of Kreol Morisien, a Low-Resourced Creole Language. In *2022 IST-Africa Conference (IST-Africa)*, pages 1–10. IEEE.

Bibi Feenaz Bhaukaurally, Mohammad Haydar Ally Didorally, and Sameerchand Pudaruth. 2012. A semi-automated lyrics generation tool for Mauritian Sega. *IAES International Journal of Artificial Intelligence*, 1(4):201–213.

Zaheenah Boodeea and Sameerchand Pudaruth. 2020. Kreol Morisien to English and English to Kreol Morisien Translation System using Attention and Transformer Model. *International Journal of Computing and Digital Systems*, 9(6):1143–1153.

Arnaud Carpooran. 2011. *Diksioner morisien : Premie diksioner kreol monoleng - 2em edision*. Sainte Croix, Mauritius: Koleksion Text Kreol.

Arnaud Carpooran, Rada Tirvassen, Nita Rughoonundun, Daniella Police-Michel, Alain Romaine, Shameem Oozeerally, Sushila Seetaram, Pascal Nadal, Ashish Beesoondiyal, Benazir Chady, and et al. 2011. *Lortograf Kreol Morisien*. Akademi Kreol Morisien, Ministry of Education and Human Resources.

Paul Choy. 2014. *Korek!: A Beginner's Guide to Mauritian Creole*. Pachworks.

Simon Clematide, Karina Frick, Noëmi Aepli, and Jean-Philippe Goldman. 2016. Crowdsourcing Swiss Dialect Transcriptions for Assessing Factors in Writing Variations. *Bochumer Linguistische Arbeitsberichte*, pages 62–67.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards Universal Dependencies for Bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Raj Dabre and Aneerav Sukhoo. 2022. MorisienMT: A Dataset for Mauritian Creole Machine Translation. *arXiv preprint arXiv:2206.02421*.

Raj Dabre, Aneerav Sukhoo, and Pushpak Bhattacharyya. 2014. Anou tradir: Experiences in building statistical machine translation systems for Mauritian languages–creole, English, French. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 82–88.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *The 60th Annual Meeting of the Association for Computational Linguistics*, pages 7676–7685. Association for Computational Linguistics (ACL).

Cheikh M Bamba Dione. 2021. Multilingual Dependency Parsing for Low-Resource African Languages: Case Studies on Bambara, Wolof, and Yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92.

Morris Football Fans. 2023. Social Media Post: Kot sa Football Lontan la? Facebook post, October 8.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2021. TuDeT: Tupían Dependency Treebank (Version v0.2).

Joshua A Fishman. 2011. *In praise of the beloved language: A comparative view of positive ethnolinguistic consciousness*, volume 76. Walter de Gruyter.

Baby Gobin-Rahimbux, Ishwaree Maudhoo, and Nuzhah Gooda Sahib. 2023. KreolStem: A hybrid language-dependent stemmer for Kreol Morisien. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–19.

Jürgen Handke. 2012. Language Typology - Language Contact. In *Virtual Linguistics Campus*. Oxford University Press.

Jimmy Desiré Harmon. 2015. *A critical ethnography of Kreol Morisien as an optional language in primary education within the Republic of Mauritius*. Ph.D. thesis, University of the Western Cape.

Lloyd Hill. 2010. Language and status: On the limits of language planning. *Stellenbosch Papers in Linguistics*, 39:41–58.

Nora Hollenstein and Noëmi Aepli. 2014. Compilation of a Swiss German dialect corpus and its application to PoS tagging. In *Proceedings of the first workshop on applying NLP tools to similar languages, varieties and dialects*, pages 85–94.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2022. A morph-based and a word-based treebank for Beja. In *TLT 2021-20th International Workshop on Treebanks and Linguistic Theories. 21-25 March 2021, Sofia, Bulgaria*, pages 48–60.

Sarah Karnes. 2023. *Automatic Parsing of Polynesian Languages: A Case Study in Cook Islands Māori*. Undergraduate Honors Thesis, Dartmouth College.

Sarah Karnes, Rolando Coto-Solano, and Sally Akevai Nicholas. 2023. Towards Universal Dependencies in Cook Islands Māori. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 124–129.

Tannon Kew. 2020. Language Representation and Modelling for Swiss German ASR.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Heather Lent, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard. 2022. What a creole wants, what a creole needs. In *LREC*.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic Speech Recognition and query by example for Creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Alice Millour and Karën Fort. 2018. Krik: First steps into crowdsourcing POS tags for Kréyòl Gwadloupéyen. In *CCURL 2018*.

Alice Millour and Karën Fort. 2020. Text corpora and the challenge of newly written languages. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for*

Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), pages 111–120.

Ludovic Mompelat, Daniel Dakota, and Sandra Kübler. 2022. How to Parse a Creole: When Martinican Creole Meets French. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4397–4406.

Nicholas Natchoo, Miven Tirvengadum, Hansinee Beeharee, Jennifer Bonne, Sandrine Cunnusamy, Jennita Dindyal, Stéphanette Ducasse, Isabelle Louise, Leveen Nowbotsing, Vedita Jokhun, and Isstiac Gooljar. 2017. *Ki Pase La? - Grad 2 Volim 2*. Mauritius Institute of Education.

Wajiha Noormamode, Baby Gobin-Rahimbux, and Mohammad Peerboccus. 2019. A speech engine for Mauritian Creole. In *Information Systems Design and Intelligent Applications: Proceedings of Fifth International Conference INDIA 2018 Volume 2*, pages 389–398. Springer.

Jessen Nullathemby. 2023. Social Media Comment on L'Express Maurice: Parlement: Hans Margueritte, CEO de la NEF avec un School Certificate. Facebook comment, October 18.

Hyunji Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for Polysynthetic Languages: A Case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142.

Aleksandar Petrov, Emanuele La Malfa, Philip HS Torr, and Adel Bibi. 2023. Language Model Tokenizers Introduce Unfairness Between Languages. *arXiv preprint arXiv:2305.15425*.

Daniella Police-Michel, Arnaud Carpooran, and Guilhem Florigny. 2012. *Gramer Kreol Morisien*, volume I Dokiman Referans. Akademi Kreol Morisien, Ministry of Education and Human Resources.

Rajini Poomen. 2023. Social Media Comment on L'Express Maurice: Parlement: Les questions au menu de mardi prochain. Facebook comment, October 18.

Sameerchand Pudaruth, Lallesh Sookun, and Arvind Kumar Ruchpaul. 2013. English to Creole and Creole to English Rule Based Machine Translation System. *International Journal of Advanced Computer Science and Applications*, 4(8).

Sameerchand Pudaruth, Aneerav Sukhoo, Somveer Kishnah, Sheeba Armoogum, Vandanah Gooria, Nirmal Kumar Betchoo, Fadil Chady, Ashminee Ramoogra, Hiteishee Hanoomanjee, and Zafar Khodabocus. 2021. Morisia: A Neural Machine Translation System to Translate between Kreol Morisien and English. *InTRAlinea: Online Translation Journal*, 23.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal dependencies for western sierra puebla nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies Treebank. *arXiv preprint arXiv:2106.03391*.

Risto Saarinen. 2016. La distribution des fonctions des langues dans un contexte multilingue : cas de l'Île Maurice. Master's thesis, University of Turku.

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: how to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Larissa Schmidt, Lucy Linder, Sandra Djambazovska, Alexandros Lazaridis, Tanja Samardžić, and Claudiu Musat. 2020. A Swiss German Dictionary: Variation in Speech and Writing. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2720–2725, Marseille, France. European Language Resources Association.

Pieter AM Seuren. 1995. Notes on the history and the syntax of Mauritian Creole. *Linguistics*, 33:531–577.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

Aneerav Sukhoo, Pushpak Bhattacharyya, and Mahen Soobron. 2014. Translation between English and Mauritian Creole: A statistical machine translation approach. In *2014 IST-Africa Conference Proceedings*, pages 1–10. IEEE.

Anand Syea. 2013. *The syntax of Mauritian creole*. A&C Black.

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the third workshop on universal dependencies (udw, syntaxfest 2019)*, pages 70–77.

Francis Tyers and Robert Henderson. 2021. A Corpus of K'iche' Annotated for Morphosyntactic Structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis M. Tyers, Mariya Sheyanova, and Jonathan North Washington. 2018. UD Annotatrix: An annotation tool for Universal Dependencies. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories (TLT16)*, pages 10–17.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-Konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161.

Irina Wagner, Andrew Cowell, and Jena D Hwang. 2016. Applying Universal Dependency to the Arapaho Language. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 171–179.

9. Language Resource References

Caron, Bernard and Strickland, Emmett and Courtin, Marine and Gerdes, Kim and Guillaume, Bruno and Kahane, Sylvain and Kennedy Ajede, Chika and Onwuegbuzia, Emeka and Tella, Samson. 2020. *UD Naija NSC*. NaijaSynCor Project. PID https://github.com/UniversalDependencies/UD_Naija-NSC.

Kahane, Sylvain and Guillaume, Bruno and Caron, Bernard and Jiang, Katharine and Davan, Marvellous. 2023. *SUD Zaar Autogramm*. Autogramm ANR project. PID https://github.com/surfacesyntacticud/SUD_Zaar-Autogramm.

Appendix 1: Hyperparameters for UDPipe 2 Training

```
batch_size: 32
beta_2: 0.99
char_dropout: 0
cle_dim: 256
clip_gradient: 2.0
dropout: 0.5
epochs: [(3, 0.001), (3, 0.0001)]
exp: None
label_smoothing: 0.03
max_sentence_len: 120
min_epoch_batches: 300
parse: 1
parser_deprel_dim: 128
parser_layers: 1
predict: False
predict_input: None
predict_output: None
rnn_cell: LSTM
rnn_cell_dim: 512
rnn_layers: 2
rnn_layers_parser: 1
rnn_layers_tagger: 0
seed: 42
single_root: 1
tag_layers: 1
threads: 4
variant_dim: 128
we_dim: 512
wembedding_model: bert-base-
multilingual-uncased-last4
word_dropout: 0.2
```