Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen

Alice Millour, Karën Fort

Krik: First Steps into Crowdsourcing POS tags for Kréyòl Gwadloupéyen

Alice Millour, Karën Fort
Sorbonne Université, STIH - EA 4509, Paris, France
alice.millour@etu.sorbonne-universite.fr, karen.fort@sorbonne-universite.fr

Abstract

This article presents the adaptation to Guadeloupean Creole of a project of crowdsourcing part-of-speech (POS) tags initially designed for a French regional language, Alsatian. We do not detail here the specifically developed crowdsourcing platform and methodology, but rather focus on the construction of the required elements for a language to be a candidate for this task: i) an open-source raw corpus, ii) a tokenizer, iii) adapted annotation guidelines, iv) a minimal reference, and, preferentially, v) one or two baseline tagger(s). After describing the preliminary work we have carried out for Guadeloupean Creole to comply with these prerequisites, we present the first results on crowdsourcing POS tags through the platform specifically developed for this task: `Krik`.

Keywords: Guadeloupean Creole, crowdsourcing, POS tagging, less-resourced languages

1. Introduction

Despite the progress made in unsupervised learning, manually annotated corpora are still necessary both to develop and to evaluate natural language processing (NLP) tools. However, building such corpora is notoriously expensive (see, for example, Böhmová et al. (2001)). For less-resourced languages, the (lack of) availability of language experts represents yet another obstacle to overcome. However, *a priori* non-expert speakers can be solicited online to share their linguistic knowledge and thus participate in the creation of resources for their language. To take advantage of this potential, we have designed a lightweight crowdsourcing platform enabling both the training of the participants to the task of part-of-speech (POS) tagging and the collaborative annotation of open-source corpora.

We led our first work on crowdsourcing POS tags on a Germanic French regional language: Alsatian (Millour and Fort, 2018).[1] The results obtained being promising, we tested the portability of our approach by adapting it on another less-resourced French regional language: Guadeloupean Creole (GC). This adaptation requires the availability of: i) a freely available raw corpus, ii) a tokenizer, iii) adapted annotation guidelines, iv) a minimal reference and, if possible, v) at least one baseline tagger for the language considered.

After presenting the existing resources for GC, we describe the five steps of the preparatory work we had to perform for GC to be a candidate language for the crowdsourcing task. Finally, we present the very preliminary results of the latter and discuss the perspectives.

2. Related Work

2.1. Guadeloupean Creole

Guadeloupean Creole (GC) is a French-based Creole spoken in the French department and archipelago of the West Indies: Guadeloupe. GC accounts for around 600,000 speakers (400,000 in Guadeloupe, and approximately 200,000 elsewhere (Colot and Ludwig, 2013)). GC is very close to the other main variety of Antillean Creole: Martinicain Creole (MC). Yet some lexical and morphological features distinguish them (see for instance the personal pronouns "*man*"/"*an*" in MC, "*moin*"/"*mwen*" in GC for the first person singular pronoun "I", or the possessive pronouns "*fidji**'w***" in MC, "*figi **a'w***" in GC ("**your** face")). What is more, GC presents a greater linguistic variation as a result of its less compact geography (Observatoire des pratiques linguistiques, 2005). Additionally, no spelling standard is recognized as the legitimate norm among speakers. Two main spelling systems coexist: one has been developed by the GEREC-F[2] (Ludwig et al., 1990), and later modified by Bernabé (2001), the other has been introduced by Hazaël-Massieux (2000). In particular, no agreement has been reached regarding the positioning towards French orthography when it can be invoked. For instance, both the forms "*chien*" (French for "dog") and "*chyen*" can be found in GC. Similarly, "*latè*" is the agglutination of the French determiner "*la*" ("the") and proper noun "*Terre*" ("Earth"). It is generally perceived as a unique entity, meaning "Earth" as a whole, and is consequently written as such. Still, we have found occurrences of the separated form, which is considered as erroneous by creolists. Generally speaking, we have encountered in our yet relatively small corpus a great variety of spelling alternatives, regardless of the conventions suggested by the two main standards. For instance, the use of the hyphen between nouns and postponed determiners (e.g. "*tifi-la*" or "*tifi la*" ("the young girl")), or the suppression of the space between adjectives and nouns in some cases (e.g. "*jenn fi*", "*jenn-fi*, "*jennfi*" ("young girl") (Delumeau, 2006)), are not consistent across the corpus.

The case of "*a pa*"/"*apa*" also exemplifies the poor penetration of the standards among the speakers. While Ludwig et al. (1990) introduced a graphic convention to distinguish "*a pa*" (negative existential) found in context such as "*A pa pas ou ni lajan* [...]"[3] ("Not because you have money [...]"), from "*apa*" ("apart (from)"), two out of three GC speakers we have been working with had never encountered the separated form.

Furthermore, GC presents a reduced inflectional and derivational morphology: the plural is indicated only by

[1] See: `http://bisame.paris-sorbonne.fr`.

[2] The GEREC-F (Groupe d'Études et de Recherches en Espace Créolophone et Francophone) is the investigation group for Creole and French speaking areas.

[3] This example was taken from (Delumeau, 2006).

the particle "*sé*" ("*timoun-la*" ("the child"), "*sé timoun-la*" ("the children")), the verbal lexeme is mainly invariable, and tenses and aspects are marked by combinations of particles. It makes it impossible to identify the part-of-speech of some words independently of the context: for instance, "*manjé*" can both mean "eat" (in its infinitive and conjugated form) and "food".

2.2. Existing Resources for GC

Some work can be found regarding GC processing. For instance, Delumeau (2006) introduces a linguistic description for GC in a natural language generation perspective, Carrión Gonzalez and Cartier (2012) detail the existing lexical resources for various French-based Creoles, Schang (2013) presents a metagrammar for GC, and Schang et al. (2017) describes the result of the annotation of coreference relations of a transcribed spoken GC corpus (the same we use here).
Yet, to our knowledge, no POS tagged corpus or tagger was available until now.

3. Methodology

3.1. Raw Corpus

To ensure the further availability of the annotated resources produced through the platform, we have focused on gathering a freely available corpus, which can be described as "opportunistic" (McEnery and Hardie, 2011), thus introducing a bias in term of content. In fact, our corpus is made of texts gathered from two sources:

- The COCOON[4] database, which contains 11 transcripts[5] of conversations led in GC (we actually used 10 out of them, for the 11^{th} contained too many French utterances), available under the CC BY-NC-SA license.[6]

- Wikipedia: we collected the proverbs found on the French page for GC[7], and the 17 articles from the Wikimedia incubator for a Guadeloupean Creole encyclopedia.[8] This corpus, C_{Wiki}, contains 74 sentences adding up to 873 tokens.

3.2. Tokenizer and Annotations Guidelines

For the sake of adaptability, we chose to work with the universal POS tagset presented in (Petrov et al., 2012), which synthesizes the tagsets of 22 languages and can be adapted to the specificities of each language.[9] Initially, the only modification we made was to have the X category ("Others", a catch-all category hard to interpret) to match only the cases of code-switching, which can not be analyzed as loan words. Eventually, just as for Alsatian, we had to further enrich this tagset with four additional categories described hereafter. The refinement of the tagset, the adaptation of the tokenizer, the elaboration of the guidelines and the building of the reference are simultaneous processes including back and forth adjustments.
The tokenizer, initially developed for Alsatian, has been provided by D. Bernhard (LiLPa, Université de Strasbourg) and adapted to the specific needs of Creole. Two kinds of operations were added to the classic tokenization process:

- Merging: decided when space-separated tokens matched a unique morphological entity. For instance, the sequence "*ki jan*" (meaning "how", literally "which kind") only appears in its separated form in our corpus. Although one could be tempted to annotate "*ki*/PRON *jan*/NOUN", this goes against the intuition of native speakers. We thus created the token "*ki_jan*/ADV". The same operation was led on the equally more intuitive "*ki_tan*/ADV" ("when", literally "which time"), "*ki_koté*/ADV" ("where", literally "which side"), etc. Prepositional locutions such as "*a fòs*" ("by dint of") were also merged for annotation consistency reasons.

- Splitting: applied when punctuation-separated tokens matched a sequence of two morphological entities understandable as such on their own. This case is exemplified by the cases of postponed determiners "*-la*" (definite article) and "*-lasa*" (demonstrative determiner), which are stick to the noun they determine in their usual form (e.g. "*Egliz-lasa*", "this Church").

Note that we did not split the tokens containing an apostrophe, indicating a contraction, but which refer to a sole interpretation for native speakers. This is the case for the tokens such as "*k'ay*/PART+VERB", contraction of "*ka*" (particle for the present tense) and "*ay*" (3^{rd} person singular for the verb "have"), for which the tokenization "*k' ay*" makes the reading and understanding confusing. For the same reason, tokens involving pronouns such as "*ba'y*/ADP+PRON" ("for him/her"), "*trapé'y*/VERB+PRON" ("catch him/her"), or "*sa'w*/PRON+PRON" (contraction of "*sa*" ("this") and "*ou*" ("you")), were not split.
These considerations resulted in the addition of 4 new categories to the universal tagset: PRON+PRON, PART+VERB, ADP+PRON and VERB+PRON.
The tagset we present here matches the needs encountered in our reference corpus. It should then not be considered as definitive, as the corpus we managed to gather is far from representative of all spelling habits and variants existing in Guadeloupe.
The annotation guidelines, inspired from the TCOF-POS (Benzitoun et al., 2012) guidelines, were developed to accompany both the expert annotators and the non-expert participants of the crowdsourcing project. For that reason, we followed the methodology set up for the crowdsourcing experiment on Alsatian and opted for a description of

[4] *COllection de COrpus Oraux Numériques* (Collection of digital oral corpora), see: https://cocoon.huma-num.fr/.

[5] See for instance: https://cocoon.huma-num.fr/exist/crdo/meta/crdo-GCF_1022. The full list of transcripts can be accessed by clicking on "Guadeloupean Creole French" in the "Langue(s)" section.

[6] See: https://creativecommons.org/licenses/by-nc-sa/3.0/.

[7] See: https://fr.wikipedia.org/wiki/Creole_guadeloupeen.

[8] See: https://incubator.wikimedia.org/wiki/Wp/gcf.

[9] See: http://universaldependencies.org/u/pos/all.html.

ADJ	ADV	ADP	ADP +PRON	AUX	CCONJ	DET	INTJ	NOUN	NUM
5%	7%	6%	0.1%	1%	2%	6%	0.1%	14%	0.2%
PART	PART +VERB	PRON	PRON +PRON	PROPN	PUNCT	SCONJ	VERB	VERB +PRON	X
7%	0.3%	17%	0.2%	3%	10%	3%	17%	0.2%	1%

Table 1: Tag distribution in the reference corpus.

the categories through illustrations in context. We enriched these lists of examples with "Watch out!" sections intended to prevent possible mix-ups and explain ambiguous cases.

3.3. Reference Corpus

We extracted 100 sentences (1,623 tokens) from both C_{Speech} and C_{Wiki} to build the reference corpus to be annotated by experts: C_{Ref}. It contains a sample of declarative, interrogative, imperative, either simple or complex, sentences of different sizes, and of direct and indirect speech.

While the sentences taken from the C_{Wiki} corpus can be immediately used for annotation purposes, we had to carry out some pre-processing on C_{Speech} to obtain grammatically correct sequences of ready to annotate tokens. In fact, the speech dysfluencies are fully transcribed as raw text. As a result, the "speech fragments" very seldom match an understandable utterance, let alone a full grammatical proposition, when taken out of context. As a consequence, we were forced to alter the original corpus in two main ways:

- cleaning up some of the dysfluencies such as the ellipses which resulted in some token being arbitrarily split. In the following example "*gwoka*" meaning literally "big drum", a Guadeloupean music genre:

  1. "*sé pou sa jodijou nou ka respékté gwo...*" ("This is why we respect big...")
  2. "*ka*" ("drum")

  In fact, and although "*gwo ka*" can be found in its space separated form[10], the first utterance is grammatically incomplete. Not to mention that the separated token "*ka*" is ambiguous and could be annotated either NOUN or PART, if presented without any context.

- bringing together the "speech fragments", such as:

  1. "*Lagwadloup dévlopé pli*" ("Guadeloupe has developed more")
  2. "*vit sé on grand tè*" ("fast, it is a big land")

We further split C_{Ref} into two groups, each annotated independently by two annotators (either a GC speaker or expert of the annotation task). The 100 sentences were then manually adjudicated. Table 1 gives the tag distribution across our 100 sentences reference corpus.

3.4. Baseline Taggers

The existing crowdsourcing platform enables participants to correct pre-annotations, thus easing and fastening the annotation process (Fort and Sagot, 2010). Two pre-annotation tools were used for this.

The first pre-annotation tool we developed relies on the weak morphological complexity of GC and uses the 100 most frequent unambiguous tokens of our corpus. This list is undoubtedly not representative of the most frequent words in GC, some common nouns being for instance repeated several times in our corpus and therefore overrepresented. Nonetheless, the most frequent words being also frequent in absolute (for instance, the particle "*ka*" represents 4.6% of the corpus, the pronoun "*an*" (meaning "I") 3.6%, the verb "*sé*" ("be") 2.8% etc.), the basic associative python script we created from this list enabled to annotate 37% of our raw corpus.

Our second pre-annocation tool is the `MElt` tagger (Denis and Sagot, 2012), used without an additional lexicon. To overcome the evaluation bias due to the very small size of our corpus, we split C_{Ref} into ten sets of two subcorpora: $C_{Training,1..10}$ (85 sentences randomly extracted from C_{Ref}) and $C_{Test,1..10}$ (containing the 15 remaining sentences). They were used respectively to train and to evaluate the tagger. We trained `MElt` on the 10 sets and obtained an average accuracy of 82%. The $MElt_{Init}$ tagger was chosen among them.

4. Krik

4.1. The Crowdsourcing Platform

The five requirements having been fulfilled, we provided the dedicated crowdsourcing platform: `Krik`[11] with the required elements. After a training phase of 4 sentences taken from C_{Ref}, which must be entirely properly annotated, the participants access the production phase in which they annotate full sentences extracted from C_{Raw}. This phase is illustrated on Figure 1. Whenever the two pre-annotation tools agree on the annotation for a given token, the consensual tag is suggested to the participant who can either validate or reject it (see on Figure 1 the case "*ou*" ("you") and the suggested tag PRON). When the pre-annotation tools disagree, the two discordant tags are suggested (see on Figure 1 the case "*la*" (postponed determiner) and the suggested tags ADP and DET). In either cases, the full list of tags is available.

4.2. Results

So far, 35 persons created an account on the platform, 17 completed the training phase, and 11 actually produced a total of 1,205 annotations during a period of 9 days. This is far from enough, both in terms of participation and of production.

Still, the annotation on `Krik` resulted in a new, freely available, collaboratively annotated corpus of 74 sentences (698 tokens).

The annotation platform does not compel the participants to annotate every token in the production phase. Thus, we filled the gaps with the $MElt_{Init}$ annotations to obtain consistent tag sequences. This resulted in a new corpus of 933 tokens: C_{Krik}. The addition of this corpus to $C_{Training}$ in the training of the tagger leads to a drop in

[10] This is not the convention used in the corpus we gathered.

[11] See: `http://krik.paris-sorbonne.fr`.

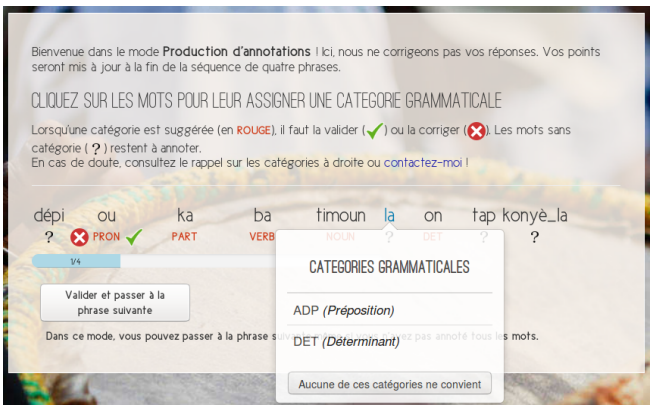

Figure 1: Screen shot of the annotation production phase on the Krik platform.

performance, even though the size of the corpus increases of 62%. This reflects the poor quality of the annotations crowdsourced so far. This is consistent with the low confidence score we calculated for the participants[12], and the difficulties they expressed.
To understand the cause of the errors we manually inspected and corrected the crowdsourced annotations. Among the 124 tokens (nearly 13% of C_{Krik}) that we corrected we identified two main difficulties:

- issues related to the nature of the raw corpus: the C_{Speech} corpus has not been entirely corrected, as described in Section 3.3. This caused the presence in C_{Raw} of unintelligible, hence discouraging, sentences.

  What is more, some additional tokenization problems have been brought to our attention. This explains the thin difference in number of tokens before and after correction. Most of these problems concerned tokens that had to be manually split, but we also encountered tokens as "*anba la*" meaning "down" which must be merged when they are not followed by a noun. We also noticed some spelling mistakes, such as the missing capital letter of "*étazini*" ("United States"), that led the participants to erroneously annotate the token as NOUN instead of PROPN.

- guidelines flaws: the guidelines we initially proposed could not prevent certain mix-ups such as the confusion for "*té*" between the verb and the particle expressing the past. Besides, the case of code-switching remains challenging, especially given the proportion of loan words in GC. During this first experiment, "French words" were alternatively annotated with either the category X or their corresponding category in French (which does not necessarily match the expected tag in GC).

Table 4.2. shows the results of the training of MElt on the corpora described above. The best results are obtained

[12]We do not detail our methodology of evaluation for the users here, for more information, see (Millour and Fort, 2018).

Training corpus	Size (tokens)	Accuracy
$C_{Training}$	1,501	82%
C_{Krik}	933	76%
C_{Krik}+$C_{Training}$	2,434	81%
$C_{KrikCorrected}$+$C_{Training}$	2,439	**84%**

Table 2: Accuracy of the trained MElt taggers.

with the manually corrected corpus $C_{KrikCorrected}$, which reaches a 84% accuracy on C_{Test}.
These results highlight two points:

- The pre-processing of the raw corpus and the annotation guidelines can and must be improved. In fact, these enhancements are compulsory as our experiment shows that a drop in the annotation quality may degrade the performance of the tagger and could remain unnoticed.

- The performance of the tagger could easily be enhanced if more annotations were to be crowdsourced. As already stated by Guillaume et al. (2016) and confirmed by our own experience on Alsatian (Millour and Fort, 2018), quality rises with participation. As a comparison, we have collected, thanks to the Alsatian platform, 18,917 annotations in 73 days, reaching a 93% accuracy for manual annotation. The annotation campaign we led resulted in a newly POS tagged corpus of 6,878 tokens. This is why efforts on advertising about the platform should be carried on.

5. Conclusion and Perspectives

We have described the steps for preparing the necessary elements for a language to benefit from the POS tags crowdsourcing platform developed in our previous work.
This process resulted in the development of new resources for GC, among which a corpus of 2,439 tokens annotated with POS tags (Millour and Fort, 2018) and the first dedicated POS tagger reaching 84% accuracy. They are both freely available under the CC BY-NC-SA license.[13]
Although the data crowdsourced so far is not satisfactory, due to the low participation, the methodology has been validated for Alsatian, and we intend to follow our efforts on advertising about the crowdsourcing platform. The code of the platform is freely available on GitHub[14] under the CeCILL v2.1 license[15], and is ready to be adapted to any language fulfilling the prerequisites we presented here.

6. Acknowledgements

We wish to thank G. Feler and A. Thibault (Sorbonne Université) for participating in the building of the reference, E. Schang (LLL, Université d'Orléans) for his advice, as well as the participants of Krik for their contribution.

[13]See: https://krik.paris-sorbonne.fr/corpora.
[14]See: https://github.com/alicemillour/Bisame.
[15]See: http://www.cecill.info/.

References

Benzitoun, C., Fort, K., and Sagot, B. (2012). TCOF-POS : un corpus libre de français parlé annoté en morphosyntaxe (TCOF-POS : A freely available pos-tagged corpus of spoken french) [in french]. In *Proc. of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 99–112, Grenoble, France, June. ATALA/AFCP.

Bernabé, J. (2001). *La graphie créole*. Guides du CAPES de Créole, Ibis Rouge edition.

Böhmová, A., Hajič, J., Hajičová, E., and Hladká, B. (2001). The prague dependency treebank: Three-level annotation scenario. In *Treebanks: Building and Using Syntactically Annotated Corpora*, pages 103–127. Kluwer Academic Publishers.

Carrión Gonzalez, P. and Cartier, E. (2012). Technological tools for dictionary and corpora building for minority languages: example of the French-based Creoles. In *Proc. of LREC'2012 (Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012))*, pages 47–53, Istanbul, Turkey, May.

Colot, S. and Ludwig, R. (2013). Guadeloupean and Martinican Creole. In Susanne Maria Michaelis, et al., editors, *The survey of pidgin and creole languages. Volume 2: Portuguese-based, Spanish-based, and French-based Languages*. Oxford University Press.

Delumeau, F. (2006). *Une description linguistique du créole guadeloupéen dans la perspective de la génération automatique d'énoncés*. Ph.D. thesis, Université de Nanterre - Paris X.

Denis, P. and Sagot, B. (2012). Coupling an Annotated Corpus and a Lexicon for State-of-the-art POS Tagging. *Language Resources and Evaluation*, 46(4):721–736.

Fort, K. and Sagot, B. (2010). Influence of pre-annotation on POS-tagged corpus development. In *The Fourth ACL Linguistic Annotation Workshop*, pages 56–63, Uppsala, Suède, July.

Guillaume, B., Fort, K., and Lefebvre, N. (2016). Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proc. of International Conference on Computational Linguistics (COLING)*, pages 3041–3052, Osaka, Japan, December.

Hazaël-Massieux, M.-C. (2000). *Ecrire en créole : Oralité et écriture aux Antilles*. L'Harmattan.

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

McEnery, T. and Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Millour, A. and Fort, K. (2018). Toward a Lightweight Solution for Less-resourced Languages: Creating a POS Tagger for Alsatian Using Voluntary Crowdsourcing. In *Proc. of Language Resources and Evaluation Conference (LREC'2018)*, Miyazaki, Japan, May.

Observatoire des pratiques linguistiques. (2005). *Les créoles à base française*, volume 5. DGLFLF.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In *Proc. of Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Schang, E., Antoine, J.-Y., and Lefeuvre-Halftermeyer, A. (2017). Les chaînes coréférentielles en créole de la Guadeloupe. In *Proc. of TALN'2017 (DILITAL workshop)*, pages 54–61, Orléans, France, June.

Schang, E. (2013). Extended Projections in a Guadeloupean TAG Grammar. In *Proc. of ESSLLI 2013 (HMGE workshop)*, pages 55–67, Düsseldorf, Germany, June.

6.1. Language Resource References

Millour, Alice and Fort, Karën. (2018). *POS Tagged Corpus of Guadeloupean Creole*.