Automatic Transcription of Grammaticality Judgements for Language Documentation

Éric Le Ferrand
Boston College
leferran@bc.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

Descriptive linguistics is a sub-field of linguistics that involves the collection and annotation of language resources to describe linguistic phenomena. The transcription of these resources is often described as a tedious task, and Automatic Speech Recognition (ASR) has frequently been employed to support this process. However, the typical research approach to ASR in documentary linguistics often only captures a subset of the field's diverse reality. In this paper, we focus specifically on one type of data known as grammaticality judgment elicitation in the context of documenting Kréyòl Gwadloupéyen. We show that only a few minutes of speech is enough to fine-tune a model originally trained in French to transcribe segments in Kréyol.

1 Introduction

Under-resourced languages, characterized by insufficient data to train common statistical or neural models, stand in contrast to high-resource languages like English, French, and Mandarin. The EGIDS scale (Lewis and Simons, 2017) offers a more nuanced classification, assessing endangerment based on socio-political factors such as the number of speakers or support from public institutions. This scale can be used to assess a language's resource level and its ability to acquire linguistic resources. For instance, the presence of media representation in a language suggests a wealth of transcribed speech, while languages lacking media exposure or educational institutions typically possess limited data, often from descriptive linguistics efforts.

In the context of under-resourced languages falling below level 4 on the EGIDS scale, where a comprehensive educational system is lacking, attention has been directed towards advancing speech technologies. These innovations aim to assist linguists in overcoming the transcription bottleneck, thereby expediting the creation of new transcribed spoken resources. One potential procedural approach involves collecting a few hours of transcribed monolingual speech in the target language, training a model with this data, and subsequently employing the model to automatically transcribe new recordings. Despite the demonstrated effectiveness of such a workflow (Shi et al., 2021; Prud'hommeaux et al., 2021), it's crucial to acknowledge that monolingual data capture only a subset of the diverse recordings compiled by linguists in the field.

Grammaticality judgments constitute a form of interview commonly carried out in a shared language of the linguist and the speaker, involving one or more linguists engaging with one or more speakers to discuss grammatical structures in the target language. This dynamic interaction is inherently multilingual, featuring spontaneous speech from various contributors. In the context of the documentation of Kréyol Gwadeloupéyen (ISO-gcf), this paper aims to investigate the efficacy of cutting-edge speech recognition architectures in transcribing such recordings, even when confronted with severely limited available data.

2 Background

2.1 Research context

Kréyòl Gwadloupéyen originated within the colonial setting through the interaction of French colonists and African enslaved individuals in the region of the French West Indies (Prudent, 1999; Chaudenson, 2004). Kréyòl gwadloupéyen serves as the main means of everyday interaction for a substantial portion of Guadeloupe's population. In contrast, French is employed for official and formal purposes Creole languages typically borrow much of their vocabulary from the colonial language (the *lexifier*), while their grammatical structure diverges considerably from that of the lexifier. For instance,

Proceedings of the The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 33–38

in the following example, we can observe the similarity between the lexicon in Kréyol and French (sait/sav, creole/kréyol, parler/palé) and the difference in constructions.

(1) a. Jan pa sav palé kréyol
Jean NEG know speak creole
'Jean doesn't speak creole'

b. Jean ne sait pas parler créole
Jean NEG know NEG speak creole
'Jean doesn't speak creole'

It's worth noting that although the phonological systems in Kréyol and French share similarities, their writing systems exhibit substantial differences. Kréyol's writing system is relatively recent and reflects the language's pronunciation, whereas French retains artifacts from historical pronunciations.

In the NLP community, there is a common assumption that data collected during fieldwork is primarily limited to monolingual recordings in the target language, and the response to this assumption is to develop ASR models for transcribing this data. Two main purposes, however, lead researchers to record data in an endangered language: documentary linguistics and descriptive linguistics. While both disciplines involve the collection of language data, the methods used to gather that data and its eventual use differ depending on the field (Himmelmann, 1998). Documentary linguistics involves the collection of any material in the target language to document it, while descriptive linguistics involves the collection of any material (in the target language or not) that can be used to describe the language.

Recordings created in linguistic fieldwork typically fall into one of the following categories: monolingual recordings, usually comprised of narratives or elicited speech; interviews conducted in either the target language or a more widely spoken language like French or English; and "linguistic confirmations" which could include translations or grammaticality judgements. The latter involves interactions in which a native speaker is queried about the validity of sentence structures. Typically, these interactions occur in the shared language, with the segment to be assessed presented in the target language, as demonstrated in the following example:

Linguist	*i pousé mwen sa*, Can we say that?
Speaker	*i pousé mwen? i pousé mwen sa*
Linguist	does it sound a little bit weird?
Speaker	Wait, is there *mwen sa* in your sentence?

Note that this is the traditional code-switching context as the code-switched segments are systematically the core of the conversation and are introduced predictably (e.g. "Can you say X?", "Does Y sound correct to you").

2.2 Related work

Code-switching can be defined as the alternation between two language systems within the same discourse. This phenomenon is particularly common in the context of language contact (for instance Bentahila and Davies, 1983; Valenti, 2014). This phenomenon is particularly difficult to manage for ASR as most ASR systems are trained to be monolingual.

Two main approaches have been explored to address code-switching in ASR. The first one consists of identifying the language segments in each language with a language identification model and then applying their respective monolingual ASR models (Ahmed and Tan, 2012; Weiner et al., 2012). While this approach has shown poor performances for intrasentential code-switching (i.e., when the change of language system occurs within the same sentence), the identification of similar languages such as French and French-based Creoles presents an additional challenge (Scherrer et al., 2023). A second approach has been to train the ASR model directly on bilingual data with a joint acoustic and language model (Imseng et al., 2011; Li et al., 2011; Bhuvanagirir and Kopparapu, 2012; Yeh et al., 2010; Sivasankaran et al., 2018). Several corpora have been released for major languages to train this kind of model, including English-Chinese (Shen et al., 2011; Li et al., 2012), English-Hindi (Dey and Fung, 2014), and French-Arabic (Amazouz et al., 2018). The existence of large populations bilingual in these particular language pairs makes the collection of data easier than for endangered languages where we usually have access to only a few hours of transcribed speech.

The emergence of fine-tuning approaches using highly multilingual models such as Wav2Vec XLSR (Conneau et al., 2021) or Whisper (Radford et al., 2023) opened new opportunities for under-resourced languages whose data is not suf-

ficient to train most state-of-the-art architectures. These new paths allowed more robust speech recognition systems for Indigenous, regional, and Creole languages (Le Ferrand et al., 2023; Macaire et al., 2022; Guillaume et al., 2022), where previous architectures would provide much higher error rates (Gupta and Boulianne, 2020b,a). Most of these previous studies approached these languages from a monolingual perspective with little space for multilingualism, code-switching, or empirical applications. While it is clear that highly multilingual models can be leveraged to transcribe under-resource languages with promising results, it is not clear if these models can be adapted to transcribe recordings in which a high-resource language contains many examples of code-switching in an under-resourced language.

In the field of documentary linguistics, the integration of ASR into the documentation workflow has been an enduring topic. Various approaches have been explored and have shown their efficiency in real-life scenarios. These approaches include the identification of spoken terms in a sparse transcription format (Le Ferrand et al., 2020; Bird, 2021) or the implementation of conventional ASR systems (Prud'hommeaux et al., 2021; Le Ferrand et al., 2023; Mitra et al., 2016).

3 Experiments

3.1 Data

As part of an NSF-funded student research program, a team of linguists went to Guadeloupe Island to document Kréyòl Gwadloupéyen in July 2022. During their trip, they were able to recruit language consultants. Most of the linguists involved in this project are English or Spanish speakers who also speak French with distinct accents, and they occasionally code-switch to English. Among the numerous recordings they collected, we have selected two that involve grammaticality judgments. The first recording features three speakers: two linguists and one native Kréyol speaker. The second recording involves two speakers: another linguist and another Kréyol speaker. Both recordings are primarily in French, with occasional interventions in English, and they contain segments in Kréyol that require verification. They focus on the same grammatical phenomena but use different examples. For instance, exploring the range of use of the preposition *pou*, the linguist in the first recording asked the validity of the sentence *i pousé sa pou mwen* ("he pushed it for me") while the linguist in the second recording used *i jèté sa pou mwen* ("he threw it for me").

The two recordings are 13 minutes and 20 minutes, respectively, with no overlapping speakers between the two recordings. The second recording is used for training and the first for testing. We will refer to this corpus as *CS* (Code-Switched) for the experiments in the next section.

To test the potential of monolingual data to transcribe grammaticality judgments, we incorporate a 70-minute-long corpus exclusively in Kréyol Gwadloupéyen. The audio recordings consist of spontaneous utterances about daily life topics from three male and three female speakers (Glaude, 2013).

3.2 Methods

For our task, we need an ASR model originally trained in French (or that includes French in the training data) and that can be fine-tuned to transcribe grammaticality judgements. Two main architectures are available: Wav2Vec (Conneau et al., 2021) and Whisper (Radford et al., 2023). For now, we use only Whisper as the Wav2Vec models available for French were not sufficiently large.

Whisper is an end-to-end encoder-decoder ASR system that relies on transformers. In a nutshell, the system takes 30s long audio segments and extracts log-Mel spectrograms. The resulting features are then passed into an encoder. The decoder is then trained to predict the corresponding transcription in an auto-regressive fashion. In other words, the transcription is generated one word at a time using the encoded input and the word previously transcribed.

We explore three configurations. The first is a traditional fine-tuning with our training set of grammaticality judgements (*CS* model). To determine whether the incorporation of monolingual data in Kréyol can boost the performance of the model, we train a second model on monolingual data in Kréyol and grammaticality judgements (*CS_mono* model) and a third model with only monolingual Kréyol data (*mono* model). Since the recordings are mostly in French, we also evaluate the model out of the box without any pretraining (*base*).

For all training, we use Whisper medium. We fine-tune it with the original hyperparamters [1] with only two changes. Because of memory limitations,

[1] https://huggingface.co/blog/fine-tune-whisper

we reduced the size of the training batch to 8 and increased the gradient accumulation to 2.

4 Results and discussion

	base	*CS*	*CS_mono*	*mono*
WER	50.98	40.85	40.53	79.78
CER	39.32	22.12	23.32	44.16

Table 1: WER and CER for the four models

.

	base	*CS*	*CS_mono*	*mono*
WER	98.96	40.15	46.58	65.66
CER	61.82	24.21	23.58	43.57

Table 2: WER and CER for the code-switched segments only.

oov rate	*CS*	*CS_mono*	*mono*
reference	62.16	35.13	47.29
predictions	14.68	3.8	7.69

Table 3: OOV rate for the three fine-tuned models.

We provide the overall Word Error Rate (WER) and Character Error Rate (CER) of all four models in Table 1. The first observation is that the initial model without fine-tuning (*base*) already performs relatively well. The reason is that most of the content of the recording is in French and we are only looking to adapt the model for code-switched segments. The model fine-tuned on only the code-switched dataset (*CS*) performs substantially better than the original model with an absolute WER decrease of 10%. These results are very encouraging considering the fact that only 20 minutes of speech has been used for the fine-tuning. The model fine-tuned with the code-switched data and the monolingual data together, *CS_mono*, does not show substantially better performances than the *CS* model. Finally fine-tuning only with monolingual data harms performance as it substantially modifies the objective of the model and transcribes everything in Kréyol.

For the second part of the evaluation we manually extract the code-switched segments from the transcriptions and perform the same evaluation on them (cf. Table 2). While the original model is not supposed to be robust at recognizing Kréyol, we notice that some segments were correctly transcribed. This can be explained by the similarity between French and Kréyol. The *CS* model performs noticeably better with about 60% of the Kréyol segments correctly transcribed. As before, the *CS_mono* model does not yield improvements over the *CS* model. Finally the monolingual model is not as robust as the *CS* model when transcribing the segments in Kréyol. This is perhaps because a language model is implicit in the auto-regressive generation of the transcription, meaning that the word order expected by the model is that of traditional Kréyol which differs substantially from the word order of the code-switched data.

The performance of the models given the minimal amount of data for fine-tuning suggests the ability of the models to infer the orthography of certain words. To explore this hypothesis, we calculated the out-of-vocabulary rate (OOV), which is the number of types in the test set that are not in the training set divided by the total number of types in the test set (see Table 3). Since most of the vocabulary in the corpus is French, we computed this rate only on the code-switched segments. Then, for each model we fine-tuned, we computed the number of OOVs correctly transcribed in the generated transcription out of all the tokens correctly transcribed (see Table 3)

The results suggest that when only the *CS* data is used for the fine-tuning, about 15% of the tokens correctly transcribed are OOVs. This number drops with the other models which suggests that the addition of monolingual data does not help during the training and using only *CS* data suffices to infer the orthography of unknown words. However, further analysis is necessary to confirm this trend.

Examples of transcriptions can be found in Table 4. We provide examples only from the original model and the *CS* model as it generally outperforms the others. A first observation is that the original model does not enforce a word order based on French but tries to provide French tokens that are close to the recordings (e.g. *je vais dire en cas content*: "I'm going to say in case happy"). For unknown words, it tries to make up a transcription based on French pronunciation rules (e.g. *kha*, *bai*, *moin* or *raï*). The *CS* model also tries to use known words when confronted with an OOV (*ti mwen*/*timoun*). Like the French model, it infers an orthography based on Kréyol pronunciation (*bo*/*ban*) or French pronunciation (e.g. *baï*/*bay*).

French model	*CS* model	Gold standard
enka raï blanc il y a deux choses	**an ka rayi blan** ah alors il y a deux choses	**an ka rayi blan** ah alors il y a deux choses
je vais dire **en cas content tu men amens**	vais dire **an ka contan ti mwen an mwen**	je vais dire **an ka conten timoun a mwen**
si je dis **un kha travail**	si je dis **an ka travayi**	si je dis **an ka travay**
pour moi et pour moi cest pas la même chose	**pour moi et ban mwen** cest pas la même	**pou mwen et ban mwen** cest pas la même chose
le verbe **bai** devient **bo** devant **moin**	le verbe **baï** devient **bo** devant **mwen**	le verbe **bay** devient **ban** devant **mwen**

Table 4: Examples of generated sentences. Code switched segments are bold

5 Conclusions

Here, we present the initial outcomes of our efforts to fine-tune a large speech recognition model for code-switching between two closely related languages: Kréyòl gwadloupéyen and its lexifier, French. Our focus is the specific scenario of grammaticality judgments, where linguists engage in conversations with native speakers in their shared language to evaluate the correctness of specific sentences in the target language.

The preliminary results illustrate that fine-tuning the Whisper model using just 20 minutes of speech substantially improves transcription quality. The refined model demonstrates increased resilience in transcribing noisy fieldwork data and accurately transcribes approximately 60% of the code-switched segments in our test set. This enhancement facilitates direct access to queries in the target language, which were consistently misinterpreted by the French model.

We present the findings of our initial experiments here, with more comprehensive details to be provided in future work. Subsequent experiments will involve comparing the Whisper model with alternative architectures, such as Wav2vec, and expanding the scope of experiments to encompass grammaticality judgments in other languages.

Acknowledgements

We would like to thank Dr. Fabiola Henri who agreed to share her data with us. We are also grateful for the work provided by the linguists on-site: Christian Jacobs, Erin Karnatz, Gideon Kortenhoven and José Pérez. Finally, we are grateful for the support of the Gwadeloupéyen speakers who agreed to be recorded and provide the grammaticality judgements used for this research. This material is based upon work supported by the National Science Foundation under Grant No. 1952568. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Basem HA Ahmed and Tien-Ping Tan. 2012. Automatic speech recognition of code switching speech using 1-best rescoring. In *2012 International Conference on Asian Language Processing*, pages 137–140. IEEE.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2018. The french-algerian code-switching triggered audio corpus (facst). In *LREC 2018 11th edition of the Language Resources and Evaluation Conference,*.

Abdelali Bentahila and Eirlys E Davies. 1983. The syntax of arabic-french code-switching. *Lingua*, 59(4):301–330.

Kiran Bhuvanagirir and Sunil Kumar Kopparapu. 2012. Mixed language speech recognition without explicit identification of language. *American Journal of Signal Processing*, 2(5):92–97.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Robert Chaudenson. 2004. La créolisation: théorie, applications, implications. *La créolisation*, pages 1–480.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. *Proceedings of Interspeech 2021*.

Anik Dey and Pascale Fung. 2014. A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Herby Glaude. 2013. Corpus créoloral. oai: crdo. vjf. cnrs. fr: crdo-gcf. *SFL Université Paris*.

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for automatic speech recognition, experiments on a fieldwork corpus of japhug (trans-himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178.

Vishwa Gupta and Gilles Boulianne. 2020a. Automatic transcription challenges for inuktitut, a low-resource polysynthetic language. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2521–2527.

Vishwa Gupta and Gilles Boulianne. 2020b. Speech transcription challenges for resource constrained indigenous language cree. 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 362–367.

Nikolaus P Himmelmann. 1998. *Documentary and descriptive linguistics*. Walter de Gruyter, Berlin/New York Berlin, New York.

David Imseng, Hervé Bourlard, Mathew Magimai Doss, and John Dines. 2011. Language dependent universal phoneme posterior estimation for mixed language speech recognition. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5012–5015. IEEE.

Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. 2023. Application of speech processes for the documentation of kréyòl gwadloupéyen. In *The Second Workshop on NLP Applications to Field Linguistics (Field Matters)*, page 17.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2020. Enabling interactive transcription in an indigenous community. In *28th International Conference on Computational Linguistics, COLING 2020*, pages 3422–3428. Association for Computational Linguistics (ACL).

M Paul Lewis and Gary F Simons. 2017. *Sustaining Language Use*. SIL International.

Ying Li, Pascale Fung, Ping Xu, and Yi Liu. 2011. Asymmetric acoustic modeling of mixed language speech. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5004–5007. IEEE.

Ying Li, Yue Yu, and Pascale Fung. 2012. A mandarin-english code-switching corpus. In *LREC*, pages 2515–2519.

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*, pages 2512–2520.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Lambert-Félix Prudent. 1999. Des baragouins à la langue antillaise. *Des Baragouins à la langue Antillaise*, pages 1–214.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2023. Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*.

Han-Ping Shen, Chung-Hsien Wu, Yan-Ting Yang, and Chun-Shan Hsu. 2011. Cecos: A chinese-english code-switching speech database. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)*, pages 120–123.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Sunit Sivasankaran, Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali, and Monojit Choudhury. 2018. Phone merging for code-switched speech recognition. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 11–19, Melbourne, Australia. Association for Computational Linguistics.

Eva Valenti. 2014. "nous autres c'est toujours bilingue anyways": Code-switching and linguistic displacement among bilingual montréal students. *American Review of Canadian Studies*, 44(3):279–292.

Jochen Weiner, Ngoc Thang Vu, Dominic Telaar, Florian Metze, Tanja Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. 2012. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Spoken Language Technologies for Under-Resourced Languages*.

Ching Feng Yeh, Chao Yu Huang, Liang Che Sun, and Lin Shan Lee. 2010. An integrated framework for transcribing mandarin-english code-mixed lectures with improved acoustic and language modeling. In *2010 7th International Symposium on Chinese Spoken Language Processing*, pages 214–219. IEEE.