Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyễn, Maxime Fily

Fine-tuning pre-trained models for Automatic Speech Recognition: experiments on a fieldwork corpus of Japhug (Trans-Himalayan family)

Séverine Guillaume[1] **Guillaume Wisniewski**[2] **Cécile Macaire**[1, 3]
Guillaume Jacques[4] **Alexis Michaud**[1] **Benjamin Galliot**[1]
Maximin Coavoux[3] **Solange Rossato**[3] **Minh-Châu Nguyên**[3] **Maxime Fily**[1, 5]
(1) LACITO, CNRS - Université Sorbonne Nouvelle - INALCO, France
(2) Université de Paris Cité, Laboratoire de Linguistique Formelle (LLF), CNRS, Paris, France
(3) LIG, CNRS - Université Grenoble Alpes - Grenoble INP - INRIA
(4) CRLAO, CNRS - École des Hautes Études en Sciences Sociales - INALCO
(5) Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France
severine.guillaume@cnrs.fr, guillaume.wisniewski@u-paris.fr,
cecile.macaire@univ-grenoble-alpes.fr, rgyalrongskad@gmail.com,
alexis.michaud@cnrs.fr, b.g01lyon@gmail.com,
{maximin.coavoux, solange.rossato}@univ-grenoble-alpes.fr,
minhchau.ntm@gmail.com, maxime.fily@gmail.com

Abstract

This is a report on results obtained in the development of speech recognition tools intended to support linguistic documentation efforts. The test case is an extensive fieldwork corpus of Japhug, an endangered language of the Trans-Himalayan (Sino-Tibetan) family. The goal is to reduce the transcription workload of field linguists. The method used is a deep learning approach based on the language-specific tuning of a generic pre-trained representation model, XLS-R, using a *Transformer* architecture. We note difficulties in implementation, in terms of learning stability. But this approach brings significant improvements nonetheless. The quality of phonemic transcription is improved over earlier experiments; and most significantly, the new approach allows for reaching the stage of automatic word recognition. Subjective evaluation of the tool by the author of the training data confirms the usefulness of this approach.

1 Introduction

The use of *Transformer*-type neural architectures to learn multilingual models of text and speech, coupled with methods for fine-tuning these generic representations, has opened up the possibility of developing tools for the many languages for which there is only a small amount of annotated data available. This approach has special appeal for linguistic documentation tasks: the development of semi-automatic or even automatic transcription and annotation methods based on a small amount of annotated data would reduce the annotation effort of field linguists and language workers, who could then focus their attention on linguistically and relationally meaningful tasks during fieldwork (Thieberger, 2017; Michaud et al., 2018; Partanen et al., 2020; Prud'hommeaux et al., 2021). In this multidisciplinary endeavour, it is clear that "linguists and Natural Language Processing (NLP) scientists may want to adjust their expectations and workflows so that both can achieve optimal results with endangered data" (Moeller, 2021).

The present work reports on our experiments using a pre-trained model of speech, XLS-R (Conneau et al., 2020), to develop a phonemic recognition system for a minority language of China: Japhug (Ethnologue language code: jya, Glottolog code: japh1234; see Jacques 2019, 2021). The transcription of recordings in a newly documented language is a key task for fieldworkers (linguists and language workers). It is also an interesting topic for the speech processing community, as it raises several challenges, epistemological as well as practical.

First of all, the amount of data available for such languages is very small: for instance, of the 197 languages in the Pangloss Collection (Michailovsky et al., 2014), which hosts audio recordings in various languages of the world (most of them endangered), only 44 corpora contain more than one hour of recordings. There is therefore a need for speech recognition methods that require as little training data as possible. In this respect, Japhug can be considered as an outlier, since there is a 32-hour transcribed corpus, freely available in the Pangloss

Collection[1] as well as from Zenodo (Galliot et al., 2021)[2] and as a Huggingface dataset.[3] The size of this corpus is one of the reasons for choosing Japhug as the test case for the present investigations: we wanted to be able to evaluate the amount of data that is necessary to obtain an automatic transcription of *good* quality — an important criterion here being the linguist's evaluation of the usefulness of the automatically generated transcript, as will be discussed again below.

Research in the field of resource-constrained Automatic Speech Recognition (ASR) has brought out "the importance of considering language-specific and corpus-specific factors and experimenting with multiple approaches when developing ASR systems for languages with limited training resources" (Morris et al., 2021, 4354). To mention two such factors:

- Endangered/little-described languages have structural features of their own, which may be widely different from those of the languages routinely taken into account in the work of the speech processing community. (It has even been argued that highly elaborate linguistic structures and typological oddities are more likely to be found in minority languages, for sociolinguistic reasons: Haudricourt 2017 [original publication: 1961]; Trudgill 2011.) For example, Japhug has a degree of morphosyntactic complexity that is particularly impressive, especially in view of its areal context (Jacques, 2021, *passim*).

- Speakers of minority languages frequently use words (or multi-word expressions, or even entire sentences) from other languages — typically the majority language of the country, or of the area (Moore, 2018; Aikhenvald, 2020). The presence of various loanwords, as well as cases of code-switching in the recordings, are a challenge for the automatic transcription of linguistic fieldwork data.

Conversely, there is one aspect in which automatic transcription tends to be easier for fieldwork data than for widely studied languages: namely, their high degree of orthographic transparency. Most endangered languages are languages transmitted through oral tradition, without a widely used writing system, and the transcriptions are usually made by linguists and language workers either in the International Phonetic Alphabet or in an orthography that is very close to the pronunciation. Thanks to this last characteristic one may realistically hope to achieve good quality transcriptions, as the system does not have to learn a complex spelling — unlike in the case of orthographies which have less straightforward correspondences between graphemes and phonemes (e.g. Uralic languages in Cyrillic orthography have a high degree of grapho-phonematic complexity, raising some technical difficulties: Gerstenberger et al., 2016).

The sections below are organized as follows: we start out, in section 2, by briefly describing the model we have used. Then we move on to presenting, in section 3, the results of a first set of experiments on phonemic transcription, which show that `XLS-R` does indeed allow us to produce very good quality transcriptions from a small corpus of annotated data, and that these transcriptions meet a need from the linguists conducting language documentation and conservation work. However, a second set of experiments described in section 4 shows that this result is difficult to reproduce, which leads us to qualify our initial optimistic conclusion concerning the technological dimension of the work.[4]

2 Fine-tuning pre-trained models

Principle The approach implemented in this work is based on the fine-tuning of a multilingual signal representation model, a method introduced in the field of speech recognition by Conneau et al. (2020) to build speech recognition models from little data. This approach is today at the core of many NLP models and is considered by many to be the most promising way to develop NLP and speech systems beyond the thirty or so languages (representing only 0.5 % of the world's linguistic diversity) for which there are large amounts of annotated data (Pires et al., 2019; Muller et al., 2021).

The proposed approach is composed of two steps. In the first step, `XLS-R`,[5] a multilingual model

[1] https://pangloss.cnrs.fr/corpus/Japhug

[2] https://doi.org/10.5281/zenodo.5521111

[3] https://huggingface.co/datasets/BenjaminGalliot/pangloss

[4] The models and all the scripts used in our experiments are freely available https://github.com/CNRS-LACITO/xlsr_for_pangloss.

[5] Note that many other pre-trained models are available, such as `hubert-large-ls960-ft` and `wav2vec2-`

trained in an unsupervised way on a corpus of 56,000 hours of recordings in 53 languages, is used to automatically build a language-independent, 'generic' representation of the signal. In a second step, this representation is used as input to a phonemic recognition system, trained on audio data that are time-aligned with a manual transcription provided by the linguist. This second step allows to learn how to match the signal representations with labels: in this case, it is essentially the labels corresponding to the phonemes.

In our experiments, we used the `XLS-R` multilingual model[6] and the HuggingFace API (Wolf et al., 2020) to use and fine-tune it. We ran the fine-tuning for 60 epochs (i.e. 60 iterations over the training data) to be assured that the fine-tuning had converged, and we kept the last model.

Using the model for phoneme prediction In order to apply the method described in the previous paragraph to the task of phoneme recognition, we simply defined a set of labels corresponding to the set of characters composing the phonemes. More precisely, the set of labels used for fine-tuning is made of the 44 characters that appear in at least one Japhug phoneme.[7] This technical choice is based on the experiments reported by Wisniewski et al. (2020) showing that the prediction of the characters composing the phonemes (instead of the phonemes as units) allows to obtain good predictions, sidestepping the task of explicitly listing the phonemes of the language (for example to specify that /t͡ʂʰ/ constitutes a single phoneme, noted by a trigraph: ʈ+ʂ+ʰ). For the sake of simplicity at an initial exploratory stage, we also removed from the manual transcriptions all the punctuation marks and the other miscellaneous symbols used by linguists in their transcriptions (symbols to note linguistic phenomena of emphasis or focus, for example).

To this set of grapho-phonemic labels is added the space, to delimit words, thereby coming a step closer to the development of a true speech recognition system for endangered languages. The addition of a special character marking the word boundaries is a novelty in our work;[8] it aims at allowing the system to recognize words directly. This avoids the need for post-processing or for a second system to segment the lattice of phonemes into words, such as the ones developed by Godard et al. (2018) and Okabe et al. (2021). To arrive at *bona fide* word recognition (and thus at full-fledged Automatic Speech Recognition), use of a language model is clearly the most efficient way to go, and this method has been successfully applied in the context of some minority/endangered languages (Partanen et al., 2020; Prud'hommeaux et al., 2021), but it should be remembered that there is huge diversity among the data sets available for endangered/low-resource languages, so that, surprising as it may seem, "no single ASR architecture outperforms all others" (Morris et al. 2021, 4354; see also Macaire et al. 2022 on two Creole languages). The use case addressed here is one in which the amount of text available is no greater than a few tens of thousands of words, i.e. an insufficient amount to train a language model according to standard workflows.

3 Evaluation on the Japhug language

In order to facilitate the reproduction of the experiments, the Japhug corpus is made available as a Huggingface dataset[9] which can be used off-the-shelf with the tools described here.

3.1 Experimental results

The quality of our system is evaluated using two classical metrics: the character error rate (CER), i.e. the edit distance between the reference and the prediction computed at the character level,[10] and the word error rate (WER), a similar metric computed at the word level. Note that what makes the use of the latter metric possible is that the systems we trained are capable of predicting word boundaries (which was not the case in previous work such as Adams et al. 2018).

Using a ten-hour corpus for fine-tuning `XLS-R`, the system obtains a CER of 7.4 % and a WER of 18.5 %. Figure 1 shows how the performances of a

`large-100k-voxpopuli`.

[6] This model is named `wav2vec2-large-xlsr-53` in Hugging Face API.

[7] This list is constructed simply by enumerating all the characters in the transcriptions and is not based on a phoneme inventory or a grapheme-to-phoneme mapping.

[8] Note that the use of a special character directly predicted by our model is only novel in the context of a low-resource/language documentation setting: it constitutes common practice in character-level ASR.

[9] `https://huggingface.co/datasets/BenjaminGalliot/pangloss`

[10] Our system is predicting a stream of characters and not of phonemes (as stated in §2, the label set is made of the characters used to write the phonemes) and the edit operations, at the heart of the CER computation, are defined directly on the characters. Computing the *phoneme error rate* in which each phonemes would be considered as an indivisible unit would weigh errors differently.

fine-tuned model evolve for training sets whose size is close to the corpora usually collected in fieldwork on endangered/minority languages. It turns out that the CER is already very low (12.5 %) for a training corpus containing two hours of annotated data.

These two results show that the proposed approach allows to obtain transcriptions of good quality, which reach the threshold at which the framework provided by the computer tool constitutes a useful starting point (preferable to the traditional method: a completely manual input). In particular, the performance is improved by 4 points compared to the results of Wisniewski et al. (2020), which were also based on a neural method of phonemic transcription, but which learned a signal representation only from the training data, without using a pre-trained model.

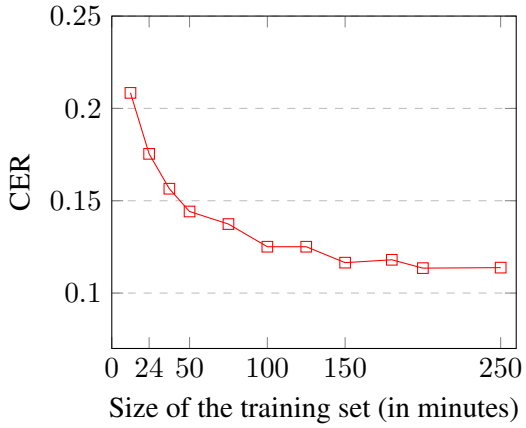

Figure 1: Evolution of performances as a function of the size of the training corpus.

The word-level error is much higher than the character-level error, but the difference is primarily due to the way in which the two evaluation metrics are defined. There are significantly fewer words than characters, so that an error at the character level (which naturally translates into an error at the word level containing it) will have a stronger impact on the WER than on the CER. A closer analysis of the results shows that our system makes few errors on word boundaries: nearly 90 % of spaces are correctly predicted.

3.2 Quality assessment of transcriptions by the linguist

To evaluate the usefulness of the system described in the previous section, a specialist of the Japhug language (Guillaume Jacques) corrected the automatic transcription of a recording that he had not yet transcribed. This pilot experiment is not systematized like that of Sperber et al. (2017) or other studies of post-editing processes in machine translation (Nitzke, 2021), and moreover concerns only 236 words, corresponding to a 2-minute recording of the Japhug language. The evaluation could therefore be dismissed as impressionistic and unreliable from the point of view of NLP tool evaluation. But it cannot be overemphasized that there is a "need for developers to directly engage with stakeholders when designing and deploying technologies for supporting language documentation" (Prud'hommeaux et al., 2021, 491). The point of view of end users is clearly significant and relevant to guide multidisciplinary team work of the type reported here.

The evaluation experiment, even though it is conducted in a way that is not standard in NLP evaluation, leads to a clear observation: the number of corrections to be made to obtain a quality transcription is much lower than the CER suggests. The linguist only had to correct 1.9 % of the characters. The figure becomes 4.2 % if punctuation is taken into account: punctuation marks are not predicted by the system – remember that they were removed from the training corpus at the preprocessing stage – and must therefore be added manually by the person taking up the automatic transcription for further processing. The corresponding WER is at 5.9 %. The difference between the estimated CER (computed on data that have been annotated beforehand) and the number of actual corrections is largely explained by the ambiguity inherent in the task of phonemic transcription: the linguist transcribing the data does not work at an exclusively phonetic-phonological level, but makes many decisions based on high-level information (in short: word identification based on context). Table 1 shows a sample of manual corrections made by the linguist to the output of our system.

The observation of a gap between the metrics and the evaluation by the user is reminiscent of similar findings obtained in the evaluation of machine translation (Wisniewski et al., 2013). Such observations are of great importance in the perspective of integrating the tools into workflows for linguistic documentation. It would seem that the actual degree of usefulness (the "real" quality) of the systems is higher than the evaluation metrics used so far would suggest. At least in the case of Japhug, the effort required to correct automatic transcrip-

①	tɕe kɯɕɯŋgɯ tɕe iɕqha @mingchao(u→ ˩)ɯraŋ nɯ tɕu pjɤŋu tɕendɤre iɕqha nɤki @yanguo kɤti rɟɤlkhɤβ ɣɯ nɯrɟɤlpu nɯ kɯ, iɕqha nɯ, iɕqha nɯ ɯftsa nɯnɯ rɟɤlpu lusɯndɤm pjɤsɯso. tɕe nɯ rɟɤlpu lusɯndɤm pjɤsɯso tɕe, tɕendɤre nɤkinɯ, sɤtɕha ra tosɤtʂoʁloʁnɯ ʑo ɕti tɕe, tɕendɤre iɕqha nɯ, @shandong nɯtɕu ɯrmi @zhangxiaobing kɯrmi ci tɯtsɣe ɯkɯβzu ci pjɤtu, tɤtɕɯ. tɕendɤre ɯrʑaβ nɯ ɯskhrɯ mɯŋɤβdi χsɯsla ma mɯtoβzu ri tɕendɤre ɯpɕi joɬoʁndʑi ɲɤphɣondʑi pjɤra matɕi sɤtɕha ra pjɤkɤtʂoʁloʁci qhe tɕe nɯra tɕetha kɯsɤɣʑi ra pɯme ma ɲɤsɯsondʑi qhe tɕe nɯ jophɣondʑi.
②	tɕe kɯɕɯŋgɯ tɕe iɕqha, @mingchao ɯraŋ nɯtɕu pjɤŋu, tɕendɤre iɕqha, nɤki, @yanguo kɤti rɟɤlkhɤβ ɣɯ, nɯrɟɤlpu nɯ kɯ, iɕqha nɯ(.→,) iɕqha nɯ, ɯftsa nɯnɯ rɟɤlpu lusɯndɤm pjɤsɯso. tɕe nɯ rɟɤlpu lusɯndɤm pjɤsɯso tɕe, tɕendɤre, nɤkinɯ, sɤtɕha ra tosɤtʂoʁloʁnɯ ʑo ɕti tɕe, tɕendɤre iɕqha nɯ, @shandong nɯtɕu, ɯrmi @zhangxiaobing kɯrmi ci, tɯtsɣe ɯkɯβzu ci pjɤtu, tɤtɕɯ. tɕendɤre ɯrʑaβ nɯ ɯskhrɯ mɯŋɤβd(er,→i) χsɯsla ma mɯtoβzu ri, tɕendɤre ɯpɕi joɬo(n→ʁ)ndʑi ɲɤphɣondʑi pjɤra matɕi, sɤtɕha ra pjɤkɤtʂoʁloʁci qhe tɕe nɯra tɕetha kɯsɤɣʑ(ɯ→i ˩)ra pɯme ma ɲɤsɯson(ɯ→dʑi) qhe tɕe nɯ jophɣondʑi.

Table 1: An excerpt from the manual corrections made to automatic transcriptions. System ①, corresponding to the setup described in §3, does not predict punctuation, nor does it predict the symbol @ (which indicates Chinese loanwords), whereas system ② predicts these two elements.

tions is considered "very low" by our expert on Japhug. A linguist's assessment of the amount of effort depends of course on many factors, including the degree of command of the target language. This makes the comparison from one case to another problematic; this is one of the difficulties encountered in interdisciplinary work between computer scientists and linguists. This point will be briefly taken up in the following paragraph.

4 Taking a critical look at the process of training statistical models

The results presented in the previous section are, to say the least, highly encouraging. They show that it is possible to achieve very good quality automatic phonemic transcriptions, even for languages for which relatively little annotated data is available (about 2 hours). Not only is the quality of the transcriptions sufficient to serve as a basis for further linguistic documentation work, but approaches based on pre-learning of representations open up the possibility of recognition at the word level, a major advance for the intended use cases (documentation of endangered languages in fieldwork). In practice, a phoneme lattice is not the best basis for further work by a field linguist. For a phoneme transcription to be complete, each individual phoneme would have to be recognizable from the audio signal, which would be contrary to all expectations, given the well-documented variability in the phonetic realization of phonemes (Niebuhr and Kohler, 2011). This variability, which carries a non-negligible part of the information contained in the signal, is particularly extensive in spontaneous speech, the object of study privileged by field linguists (Bouquiaux and Thomas, 1971; Newman and Ratliff, 2001). Thus, the basic unit for the constitution of corpora of rare languages is clearly not the phoneme, but the morpheme (and the higher-level units: word, sentence...).

Our initial results led us to consider more complex transcription tasks in which the system must also predict punctuation, as well as Chinese loanwords (cases of code-switching with the national language) found in Japhug documents (where they are transcribed according to the romanization conventions of standard Mandarin). The goal is, as before, to reduce the annotation effort of field linguists. Taking punctuation and loanwords into account essentially involves changing the preprocessing performed on the transcriptions before training.

The difficulties which we encountered during the development of this new system led us to study in a systematic way the degree of *stability* of the learning process. Neural network training is a difficult task in that it involves a very large number of parameters and relies on the optimization of a non-convex objective function. In practice, the optimization methods at the heart of deep learning rely on a very large number of hyper-parameters,[11]

[11]Hyper-parameters are special parameters the optimal

the choice of which has a direct impact on the performance of the resulting system. Thus, for the task of fine-tuning the `XLS-R` model (used in the work reported here), it is possible to change the value of more than twenty parameters that include the initial value of the learning step, its scheduling, the optimization method, the size of the batches, as well as various parameters for dropout.

We have represented in Figure 2 the performances (evaluated by the CER) obtained on the validation set during the different trainings we have performed during the development of these systems. Note that the systems were fine-tuned on a three-hour corpus (10% of which, making up 18 minutes, were used as a validation set) in order to keep the training times to a reasonable duration. The experiments we conducted with a larger corpus did not lead to improvements in the results obtained. These learning curves were obtained by varying the various parameters for optimization (training step, values for dropout, choice of the training set), but also by trying various experimental conditions: in particular, by taking into account the punctuation or not.

Among the 91 training curves shown in Figure 2, the CERs obtained on the validation set vary between 8.8 % and 28.8 % ($M = 14.8$, $S = 2.2$). Most of the learned systems perform significantly worse than the system described in our first experiments: only 6 systems have a CER at validation that is below 12.0 %, and none of them reaches the performance of the system described in section 3. Although not all of these error rates are directly comparable, these results show not only that performance on the validation set is highly sensitive to the choice of hyper-parameters (as expected), but more importantly, that the optimal value of these parameters varies across corpora, train-test splits and configurations.

However, as the results in Table 2 show, if we apply the different models obtained to the corrected text of section 3.2, the quality of the transcriptions is such that it requires only a small number of corrections. This result is all the more remarkable since these systems were only learned on 3 hours of annotated data, a reasonable amount of data to expect in scenarios of language documentation. Above all, it appears that the performance of the models on the validation set does not seem to be a reliable indicator of their quality in practice. This makes their selection and more generally their development very difficult.

	①	②	③
CER validation		8.8 %	13.9 %
WER	5.9 %	19.5 %	21.6 %
CER	4.2 %	9.1 %	6.7 %
⊖ punctuation	1.9 %	6.8 %	4.5 %
⊖ Pinyin	0.7 %	2.9 %	4.0 %

Table 2: Detailed evaluation of the various systems for phonemic transcription: ① is the system described in section 3, ② and ③ are two of the systems from our second series of experiments (described in §4): ② is the systeme with lowest CER on the validation set, and ③ that with lowest CER on the test set. These last two systems predict punctuation and the @ symbol for loanwords.

In a more qualitative way, we have reported in Table 1 an extract of the transcription of this text by the system described in section 3 and by a system predicting the punctuation. It appears that, while the first system is able to achieve a perfect transcription except for Chinese words (romanized into *Pinyin*) and punctuation marks, the second system presents properties that may be quite interesting for innovative workflows for computational documentation of languages. First of all, it places the utterance boundaries (materialized by the dot) without errors. The division into sentences constitutes a fundamental dimension of the structure of linguistic documents, and an important dimension of the work curating transcriptions for electronic publication in language archives. Moreover, the model recognizes Chinese borrowings remarkably well, paving the way for their automatic identification. Such additional treatments down the line are key to a workflow that makes the most of a range of NLP tools. The ultimate aim is to arrive at Interlinear Glossed Texts (IGT), with annotation down to the level of the morpheme; in turn, IGT corpora have considerable usefulness in research, including possibilities for automatically inferring linguistic patterns from the glossed corpora (Zamaraeva et al., 2019).

5 Conclusion

In this work, we have described how the fine-tuning of a multilingual model could be used to learn an

value of which can only be found by trial-and-error and training a system completely. Tuning hyper-parameters tends to be highly time-consuming and resource-intensive.

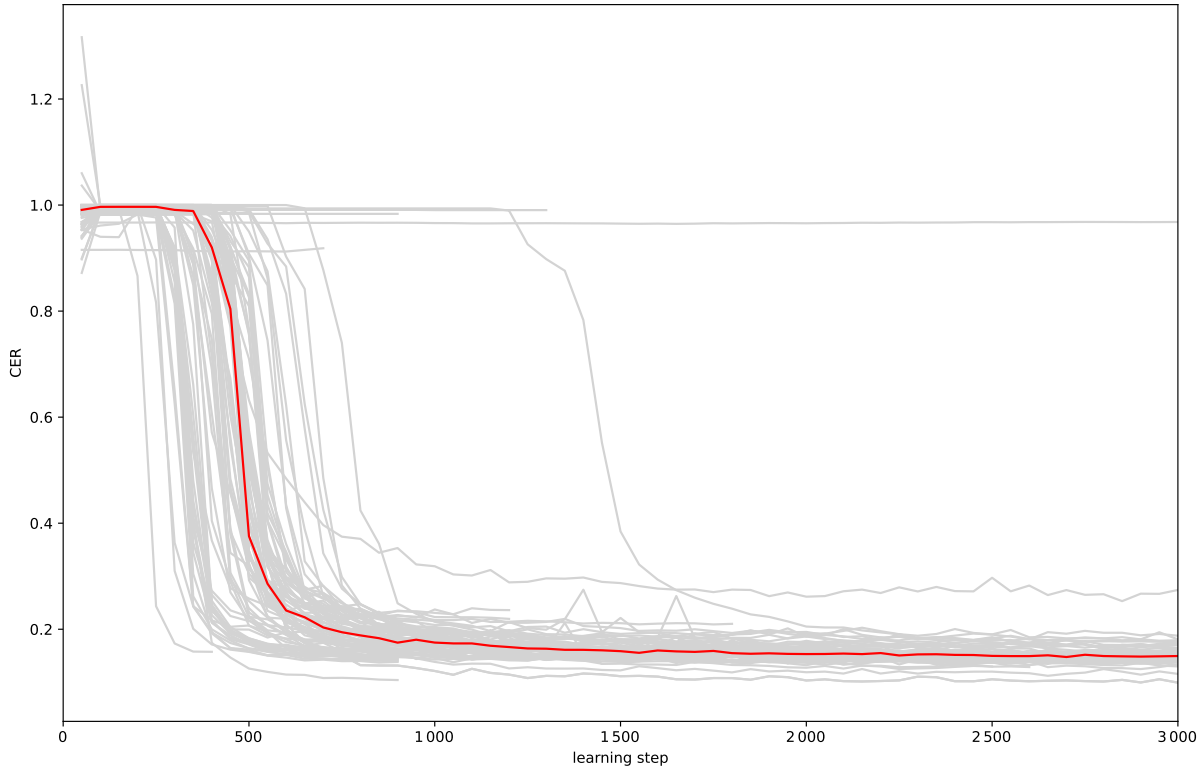

Figure 2: CER over the validation set in the course of various optimizations. The curve in red corresponds to the median value for CER at each stage.

automatic phonemic transcription system for an endangered language, and thus reduce the annotation effort of field linguists. Despite the large variability of the scores obtained on a validation set, we succeeded in developing systems whose predictions required only a small number of manual corrections by the linguist: a number that is much smaller than that estimated by the Character Error Rate (CER). This work shows the interest of this type of approach, and opens many perspectives. In particular, the approach seems to us to call for an extension of the experiments to other endangered languages (e.g. from other corpora hosted in archives of endangered languages, about which see Berez-Kroeker and Henke 2018), in order to evaluate more widely its usefulness for language documentation. We also wish, in our future work, to improve the quality of predictions at the word level, for example by integrating a language model.

Acknowledgments

We wish to express our deepest gratitude to the main Japhug language consultant, Tshendzin.

Financial support was given by *Agence Nationale de la Recherche* as part of grants ANR-10-LABX-0083 (*Laboratoire d'excellence* "Empirical Foundations of Linguistics", 2011-2024) and ANR-19-CE38-0015 ("Computational Language Documentation by 2025", 2019-2024). Financial support was also contributed by the Institute for Language Diversity and Heritage (ILARA-EPHE).

An important part of the linguistic resources used in the present work was collected in the course of the project "Himalayan Corpora: Parallel corpora in languages of the Greater Himalayan area" (ANR-12-CORP-0006).

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Alexandra Aikhenvald. 2020. Language contact and endangered languages. *The Oxford handbook of language contact*, pages 241–260.

Andrea L. Berez-Kroeker and Ryan E. Henke. 2018. Language archiving. In Kenneth L. Rehg and Lyle Campbell, editors, *The Oxford handbook of endangered languages*, pages 433–457. Oxford University Press, Oxford.

Luc Bouquiaux and Jacqueline Thomas. 1971. *Enquête et description des langues à tradition orale. Volume I : l'enquête de terrain et l'analyse grammaticale*, 1976 (2e) edition. Société d'études linguistiques et anthropologiques de France, Paris. 3 volumes.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *CoRR*, abs/2006.13979.

Benjamin Galliot, Guillaume Wisniewski, Séverine Guillaume, Laurent Besacier, Guillaume Jacques, Alexis Michaud, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2021. Deux corpus audio transcrits de langues rares (japhug et na) normalisés en vue d'expériences en traitement du signal. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Grenoble.

Ciprian Gerstenberger, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2016. Utilizing language technology in the documentation of endangered Uralic languages. *Northern European Journal of Language Technology*, 4:29–47.

Pierre Godard, Marcely Zanon Boito, Lucas Ondel, Alexandre Berard, François Yvon, Aline Villavicencio, and Laurent Besacier. 2018. Unsupervised word segmentation from speech with attention. In *Interspeech 2018*, Hyderabad, India.

André-Georges Haudricourt. 2017 [original publication: 1961]. Number of phonemes and number of speakers [translation of: *Richesse en phonèmes et richesse en locuteurs*]. *L'Homme*, 1(1):5–10.

Guillaume Jacques. 2019. Japhug. *Journal of the International Phonetic Association*, 49(3):427–450.

Guillaume Jacques. 2021. *A grammar of Japhug*. Number 1 in Comprehensive Grammar Library. Language Science Press, Berlin.

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*, Dublin, Ireland.

Boyd Michailovsky, Martine Mazaudon, Alexis Michaud, Séverine Guillaume, Alexandre François, and Evangelia Adamou. 2014. Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation*, 8:119–135.

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, 12:393–429.

Sarah Moeller. 2021. *Integrating machine learning into language documentation and description*. Ph.D. thesis, University of Colorado at Boulder.

Patrick Moore. 2018. Re-valuing code-switching: lessons from Kaska narrative performances. In *Activating the heart: Storytelling, knowledge sharing, and relationship*, Waterloo, Canada. Wilfrid Laurier University Press.

Ethan Morris, Robbie Jimerson, and Emily Prud'hommeaux. 2021. One size does not fit all in resource-constrained ASR. In *Interspeech 2021*, pages 4354–4358. ISCA.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Paul Newman and Martha Ratliff. 2001. *Linguistic fieldwork*. Cambridge University Press, Cambridge.

Oliver Niebuhr and Klaus J. Kohler. 2011. Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, 39(3):319–329.

Silvia Nitzke, Jeanand Hansen-Schirra. 2021. *A short guide to post-editing*. Number 16 in Translation and Multilingual Natural Language Processing. Language Science Press, Berlin.

Shu Okabe, François Yvon, and Laurent Besacier. 2021. Segmentation en mots faiblement supervisée pour la documentation automatique des langues. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Grenoble.

Niko Partanen, Mika Hämäläinen, and Tiina Klooster. 2020. Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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.

Matthias Sperber, Graham Neubig, Jan Niehues, Satoshi Nakamura, and Alex Waibel. 2017. Transcribing against time. *Speech Communication*, 93:20–30.

Nick Thieberger. 2017. LD&C possibilities for the next decade. *Language Documentation and Conservation*, 11:1–4.

Peter Trudgill. 2011. *Sociolinguistic typology: social determinants of linguistic complexity*. Oxford University Press, Oxford.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, pages 306–315, Marseille, France. European Language Resources Association (ELRA).

Guillaume Wisniewski, Anil Kumar Singh, Natalia Segal, and François Yvon. 2013. Design and analysis of a large corpus of post-edited translations: Quality estimation, failure analysis and the variability of post-edition. In *Proceedings of Machine Translation Summit XIV: Papers*, Nice, France.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Olga Zamaraeva, Kristen Howell, and Emily M. Bender. 2019. Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 28–38, Honolulu. Association for Computational Linguistics.