Les modèles pré-entraînés à l'épreuve des langues rares : expériences de reconnaissance de mots sur la langue japhug (sino-tibétain)

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

Les modèles pré-entraînés à l'épreuve des langues rares : expériences de reconnaissance de mots sur la langue japhug (sino-tibétain)

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) Langues et Civilisations à Tradition Orale (LACITO), Unité Mixte de Recherche 7107 CNRS - Université Sorbonne Nouvelle - Institut National des Langues et Civilisations Orientales (INALCO)
(2) Université de Paris Cité, Laboratoire de Linguistique Formelle (LLF), CNRS, 75 013 Paris, France
(3) Laboratoire d'Informatique de Grenoble (LIG), Unité Mixte de Recherche 5217 CNRS - Université Grenoble Alpes - Grenoble INP - Institut national de recherche en informatique et en automatique (INRIA)
(4) Centre de Recherches Linguistiques sur l'Asie Orientale (CRLAO), Unité Mixte de Recherche 8563 CNRS - École des Hautes Études en Sciences Sociales - Institut National des Langues et Civilisations Orientales
(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@univ-grenoble-alpes.fr, Solange.Rossato@univ-grenoble-alpes.fr, minhchau.ntm@gmail.com, maxime.fily@gmail.com

Résumé

Nous décrivons dans ce travail des résultats obtenus dans le cadre d'explorations interdisciplinaires visant à venir en appui aux linguistes « de terrain » au moyen d'outils de Reconnaissance Automatique de la Parole. Spécifiquement, nous nous focalisons sur le développement d'un système de reconnaissance de la parole pour le japhug, langue rare de Chine. L'objectif consiste à réduire l'effort de transcription des linguistes « de terrain ». Nous montrons comment une nouvelle approche neuronale fondée sur la spécialisation d'un modèle de représentation générique pré-entraîné multilingue XLS-R reposant sur une architecture de type *Transformer* permet d'améliorer significativement la qualité de la transcription phonémique dans le cas où seules quelques heures de données annotées sont disponibles, et surtout de progresser jusqu'à la reconnaissance automatique de mots. Nous relevons néanmoins des difficultés de mise en œuvre, en termes de stabilité de l'apprentissage. La question de l'évaluation de l'outil par les linguistes de terrain est également abordée.

Abstract

Testing pre-trained models on un(der-)described languages : Automatic Speech Recognition experiments on the Japhug language

We describe in this work the latest results obtained in interdisciplinary work to support "fundamental language documentation" through the use of speech recognition tools. Specifically, the focus is on the development of a speech recognition system for Japhug, an endangered minority language of China. The practical goal is to reduce the transcription workload of field linguists. We show how a new deep learning approach based on the language-specific tuning of a generic pre-trained representa-

tion model, XLS-R, using a *Transformer* architecture, significantly improves the quality of phonemic transcription, in a setting where only a few hours of annotated data are available. Most significantly, this method allows for reaching the stage of automatic word recognition. Nevertheless, we note difficulties in implementation, in terms of learning stability. The question of the evaluation of the tool by field linguists is also addressed.

MOTS-CLÉS : documentation linguistique assistée par ordinateur, reconnaissance automatique de la parole, modèles neuronaux, science ouverte, linguistique de terrain.

KEYWORDS : Computational Language Documentation, Automatic Speech Recognition, Open Science, Deep Learning, linguistic fieldwork.

1 Introduction

L'utilisation d'architectures neuronales de type *Transformer* pour apprendre des modèles multilingues du texte et de la parole, couplée à des méthodes de *spécialisation (fine-tuning)* de ces représentations génériques, a ouvert la possibilité de développer des approches de traitement automatique pour de nombreuses langues pour lesquelles il n'existe que peu de données annotées. Cette approche est particulièrement intéressante pour les tâches d'aide à la documentation linguistique : le développement de méthodes de transcription et annotation semi-automatique, voire automatique, à partir d'une petite quantité de données annotées, permettrait en effet de réduire l'effort d'annotation des linguistes de terrain ; ces dernier·ère·s pourraient alors concentrer leur attention sur des tâches significatives au plan linguistique et au plan humain (Michaud *et al.*, 2018 ; Partanen *et al.*, 2020 ; Prud'hommeaux *et al.*, 2021 ; Morris *et al.*, 2021).

Ce travail décrit notre expérience de l'utilisation d'un modèle pré-entraîné de la parole, XLS-R, pour développer un système de reconnaissance phonémique pour une langue minoritaire de Chine : le japhug (Jacques, 2019). Tâche d'une importance centrale dans le travail des linguistes de terrain, la transcription d'une langue rare soulève en outre plusieurs défis aussi bien épistémologiques que pratiques pour la communauté du traitement de la parole. Tout d'abord, la quantité de données disponible est très faible : à titre d'exemple, parmi les 197 langues représentées dans la collection Pangloss (Michaud *et al.*, 2016), qui rassemble des enregistrements sonores en diverses langues du monde (la plupart menacées d'extinction), seules 44 y disposent d'un corpus qui compte plus d'une heure d'enregistrements. Il est donc souhaitable de concevoir des méthodes de reconnaissance de la parole particulièrement peu gourmandes en données d'apprentissage.

En termes de taille, le corpus japhug fait figure d'exception puisqu'il existe un corpus transcrit de 32 heures, librement disponible dans la collection Pangloss mais aussi dans Zenodo (Galliot *et al.*, 2021). La taille de ce corpus est une des raisons du choix du japhug comme langue d'étude : nous souhaitions pouvoir évaluer la quantité de données nécessaires pour obtenir une transcription automatique de bonne qualité.

Il faut également noter que la plupart des langues rares possèdent des caractéristiques structurelles nettement différentes de celles des langues couramment abordées dans les travaux de la communauté. Par exemple, le japhug présente un degré de complexité morphosyntaxique particulièrement impressionnant au vu de son contexte aréal (Jacques, 2021). Une autre difficulté pour la transcription automatique des langues rares est la présence de nombreux emprunts lexicaux dans les enregistrements : les locuteur·trice·s utilisent en effet fréquemment des mots d'autres langues de la région,

notamment les langues nationales (Moore, 2018 ; Aikhenvald, 2020).

À l'inverse, un élément facilitateur pour la transcription automatique est que les langues à tradition orale, sans système d'écriture largement usité, sont généralement notées soit en alphabet phonétique international, soit dans une orthographe de type phonémique (à haut degré de transparence graphophonématique).

La section 2 décrit rapidement le modèle que nous avons utilisé. La section 3 expose les résultats d'une première série d'expériences, qui montrent que XLS-R permet de réaliser des transcriptions phonémiques de très bonne qualité à partir d'un petit corpus de données annotées. Une deuxième série d'expériences décrite en section 4 amène néanmoins à nuancer cette conclusion, du fait de difficultés en termes de reproductibilité des résultats.

2 Spécialisation de modèles pré-entraînés

Principe L'approche mise en œuvre ici repose sur la spécialisation d'un modèle de représentation multilingue du signal. Cette méthode, proposée par Conneau *et al.* (2020) afin de développer des modèles de reconnaissance de la parole à partir de peu de données, est aujourd'hui au cœur de nombreux modèles de TAL (Muller *et al.*, 2021).

L'approche proposée est composée de deux étapes. Dans une première étape, `XLSR-53`, un modèle multilingue entraîné de manière non supervisée sur un corpus regroupant 56 000 heures d'enregistrements en 53 langues, est utilisé pour construire automatiquement une représentation du signal. Dans une seconde étape, ces représentations sont utilisées en entrée d'un système de reconnaissance phonémique, entraîné à partir de données associées à une transcription manuelle fournie par la ou le linguiste. Cette étape permet d'apprendre à mettre en correspondance les représentations du signal avec les étiquettes employées dans la transcription phonémique.

Utilisation pour la prédiction de phonème Le jeu d'étiquettes employé correspond à l'ensemble des caractères composant les phonèmes. Nos expériences passées (Wisniewski *et al.*, 2020) ont en effet montré que la prédiction des caractères composant les phonèmes (et non pas directement des phonèmes) permettait d'obtenir de bonnes prédictions tout en faisant l'économie de l'étape qui consiste à lister explicitement les phonèmes de la langue (par exemple pour préciser que /t͡ʂʰ/ constitue un unique phonème, noté par un trigraphe : t͡+ʂ+ʰ).

À ce jeu d'étiquettes grapho-phonémiques s'ajoute l'espace, pour délimiter les mots, et par là, s'approcher un peu plus du développement d'un véritable système de reconnaissance de la parole pour les langues rares. L'ajout d'un caractère spécial marquant les frontières de mots est une nouveauté de ce travail et a pour objectif de permettre au système de reconnaître directement les mots : dans les systèmes de reconnaissance phonémique de l'état de l'art, le jeu d'étiquettes prédit par le système est composé uniquement des phonèmes de la langue (Adams *et al.*, 2018) (ou parfois des caractères composant ceux-ci : voir Wisniewski *et al.* 2020) et le système prédit un flux continu de phonèmes sans chercher à identifier les frontières de mots (celle-ci sont supprimées lors de l'apprentissage). En ajoutant l'espace au jeu d'étiquettes, le système apprend à prédire directement les frontières de mots et on fait l'économie d'un post-traitement ou d'un second système chargé de segmenter le treillis de phonèmes en mots (Godard *et al.*, 2018 ; Okabe *et al.*, 2021).

3 Évaluation sur le japhug

3.1 Résultats expérimentaux

La qualité de notre système est évaluée en utilisant deux métriques classiques : le taux d'erreur sur les caractères (CER), c'est-à-dire la distance d'édition entre la référence et la prédiction calculée au niveau des caractères, et le taux d'erreur sur les mots (WER), une métrique similaire calculée au niveau des mots. Notons que ce qui rend possible l'utilisation de cette dernière métrique est que les systèmes que nous avons entraînés sont capables de prédire les frontières de mots (ce qui n'était pas le cas dans nos précédents travaux). Tous les résultats présentés dans cette section sont évalués sur un corpus de test de 20 minutes composé de phrases choisies aléatoirement parmi toutes les données transcrites à notre disposition. (Le corpus de validation est construit de la même manière.)

En utilisant en apprentissage un corpus de dix heures, le système obtient un CER de 7,4 % et un WER de 18,5 %, résultat nettement meilleur que tous ceux obtenus jusqu'à présent pour le japhug. La figure 1 montre comment ces performances (en test, après convergence sur un ensemble de validation) évoluent pour des ensembles d'apprentissage dont la taille se rapproche des corpus usuellement collectés pour documenter une langue rare [1]. Il apparaît que le CER est déjà très faible (12,5 %) pour un corpus d'apprentissage contenant deux heures de données annotées, une taille de corpus relativement courante dans les collections des linguistes de terrain.

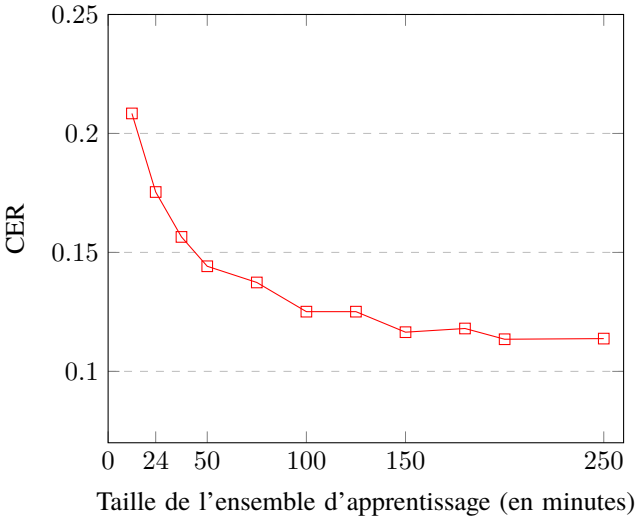

FIGURE 1 – Évolution des performances en fonction de la taille du corpus d'apprentissage.

Ces deux résultats montrent que l'approche proposée permet d'obtenir des transcriptions de bonne qualité. Les performances sont notamment améliorées de 4 points par rapport à nos premiers résultats (Wisniewski *et al.*, 2020), lesquels reposaient également sur une méthode neuronale de transcription phonémique, mais qui apprenait une représentation du signal uniquement à partir des données d'apprentissage, sans utiliser un modèle pré-entraîné.

1. Les expériences préliminaires que nous avons menées sur des corpus de plus grande taille n'ont pas montré d'amélioration significative au-delà des 10 heures utilisées dans la première expérience.

Si l'erreur au niveau des mots est bien plus élevée que l'erreur au niveau des caractères, la différence tient essentiellement à la manière dont les deux mesures d'évaluation sont définies. Il y a nettement moins de mots que de caractères, de sorte qu'une erreur au niveau d'un caractère (qui se traduit naturellement par une erreur au niveau du mot le contenant) aura un impact plus fort sur le WER que sur le CER. Une analyse plus fine des résultats montre que nos systèmes ne font que peu d'erreurs sur les frontières de mots : près de 90 % des espaces sont correctement prédits.

3.2 Évaluation « manuelle » de la qualité des transcriptions

Au-delà des évaluations quantitatives, une question importante consiste à déterminer si le canevas fourni par l'outil informatique constitue un point de départ utile : préférable à la méthode traditionnelle (saisie intégralement manuelle). Afin d'évaluer l'utilité du système, le linguiste spécialiste de la langue (Guillaume Jacques) a corrigé la transcription automatique d'un enregistrement qu'il n'avait pas encore transcrit. Cette expérience-pilote n'est pas systématisée comme celle de Sperber *et al.* (2017) ou d'autres études des processus de *post-édition* en traduction automatique (Nitzke, 2021), et ne concerne que 236 mots, correspondant à un enregistrement de 2 minutes de langue japhug. Elle débouche néanmoins sur une observation claire : le nombre de corrections à effectuer pour obtenir une transcription de qualité est beaucoup plus faible que ne le suggère le CER. Le linguiste n'a eu à corriger que 1,9 % des caractères (4,2 % en prenant en compte la ponctuation, qui n'est pas prédite par le système et doit donc être ajoutée par le correcteur), ce qui correspond à un WER de 5,9 %. Le tableau 1 montre un échantillon de corrections apportées manuellement par le linguiste à la sortie de notre système.

①	tɕe kɯɕɯŋgɯ tɕe iɕqha @mingchao(u→␣)ɯraŋg 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 mɯɲɤβdi χsɯsla ma mɯtoβzu ri tɕendɤre ɯpɕi jołoʁndʑi 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, mɯɲɤβd(er,→i) χsɯsla, ma mɯtoβzu ri, tɕendɤre ɯpɕi joło(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.

Tableau 1 – Extrait des corrections de deux transcriptions automatiques. Le système ①, correspondant au système décrit au §3, ne prédit pas la ponctuation (ni le symbole @ identifiant les emprunts chinois) ; le système ②, au contraire, prédit ces deux éléments.

Cette observation rappelle des résultats similaires obtenus dans l'évaluation de la traduction automatique, qui sont d'une grande importance dans la perspective d'une intégration des outils au sein de chaînes de traitements pour la documentation linguistique. Il semblerait que la qualité « réelle » des

systèmes soit plus élevée, en pratique, que ne le suggéraient les métriques d'évaluation employées jusqu'ici. Au moins dans le cas du japhug, l'effort demandé pour corriger des transcriptions automatiques est considéré comme « très faible » par le linguiste spécialiste de la langue japhug (Guillaume Jacques). L'appréciation portée à ce sujet par un(e) linguiste dépend évidemment de multiples facteurs, dont le degré de maîtrise de la langue qu'il ou elle possède. Cela rend la comparaison d'un cas à l'autre problématique ; c'est là une des difficultés rencontrées dans le travail interdisciplinaire entre TAListes et linguistes. Ce point sera brièvement repris dans le paragraphe qui suit.

4 Un regard critique sur le processus d'entraînement

Les résultats présentés dans la section précédente sont plus que très encourageants : ils montrent qu'il est possible de réaliser des transcriptions phonémiques automatiques de très bonne qualité, même pour des langues rares pour lesquelles relativement peu de données annotées sont disponibles. Non seulement la qualité des transcription est suffisante pour servir de base à la suite du travail de documentation linguistique, mais les approches reposant sur un pré-apprentissage des représentations ouvrent la possibilité de faire de la reconnaissance au niveau des mots, une avancé majeure pour les langues rares. En effet, la reconnaissance de phonèmes est certes une tâche intéressante pour la recherche en phonétique (voir par exemple Michaud *et al.* 2020), mais en pratique, un treillis de phonèmes n'est clairement pas la meilleure base pour la poursuite du travail par un·e linguiste de terrain. Pour qu'une transcription phonémique soit exhaustive, il faudrait que tous les phonèmes puissent être reconnus dans le signal, ce qui serait contraire à toutes les attentes, au vu de la variabilité dans la réalisation phonétique des phonèmes (Niebuhr & Kohler, 2011). Cette variabilité, porteuse d'une part non négligeable de l'information contenue dans le signal, est particulièrement grande dans la parole spontanée, objet d'étude privilégié des linguistes de terrain (Bouquiaux & Thomas, 1971 ; Newman & Ratliff, 2001). L'unité de base pour la constitution de corpus de langues rares annotés dans les règles de l'art n'est pas le phonème, mais le morphème (puis les unités de niveau supérieur : mot, phrase...).

Ces premiers résultats, obtenus lors d'un stage de master (Macaire, 2021 ; Macaire *et al.*, 2021), nous ont poussés à envisager des tâches de transcription plus complexes dans lesquelles le système doit également prédire la ponctuation, ainsi que les mots d'emprunt au chinois (langue nationale) présents dans les documents en japhug (où ils sont transcrits d'après les conventions de la romanisation *pinyin* du mandarin standard), toujours dans l'optique de réduire l'effort d'annotation des linguistes de terrain. La prise en compte de ces deux éléments consiste essentiellement à changer les pré-traitements réalisés sur les transcriptions avant l'apprentissage.

Les difficultés que nous avons rencontrées lors du développement de ce système nous ont amenés à essayer de reproduire les résultats décrits en section 3 puis à étudier de manière plus systématique la *stabilité* de l'apprentissage. En effet, l'apprentissage de réseau de neurones est une tâche réputée difficile car elle met en jeu un très grand nombre de paramètres et repose sur l'optimisation d'une fonction objectif non convexe. En pratique, les méthodes d'optimisation au cœur de l'apprentissage reposent sur un très grand nombre d'hyper-paramètres, dont le choix a un impact direct sur les performances du système obtenu. Ainsi, pour la tâche de spécialisation du modèle XLS-R que nous utilisons dans ce travail, il est possible de changer la valeur de plus d'une vingtaine de paramètres [2].

2. Pour ne citer que les principaux : valeur initiale du pas d'apprentissage, *scheduling* de celui-ci, méthode d'optimisation, taille des *batchs*, ainsi que divers paramètres du *dropout*.

Nous avons représenté en figure 2 les performances (évaluées par le CER) obtenues sur l'ensemble de validation au cours des différents apprentissages que nous avons réalisés lors de la mise au point de ces systèmes [3]. Il faut noter que les systèmes ont été appris sur un corpus de trois heures afin de garder des temps d'apprentissage « raisonnables ». Les expériences que nous avons menées avec des corpus d'apprentissage plus importants n'ont pas permis d'améliorer les résultats obtenus. Ces courbes d'apprentissage ont été obtenues en faisant varier les divers paramètres de l'optimisation (pas d'apprentissage, valeurs des différents *dropout*, choix de l'ensemble d'apprentissage), mais également en faisant varier différentes conditions expérimentales (notamment en prenant en compte ou non la ponctuation).

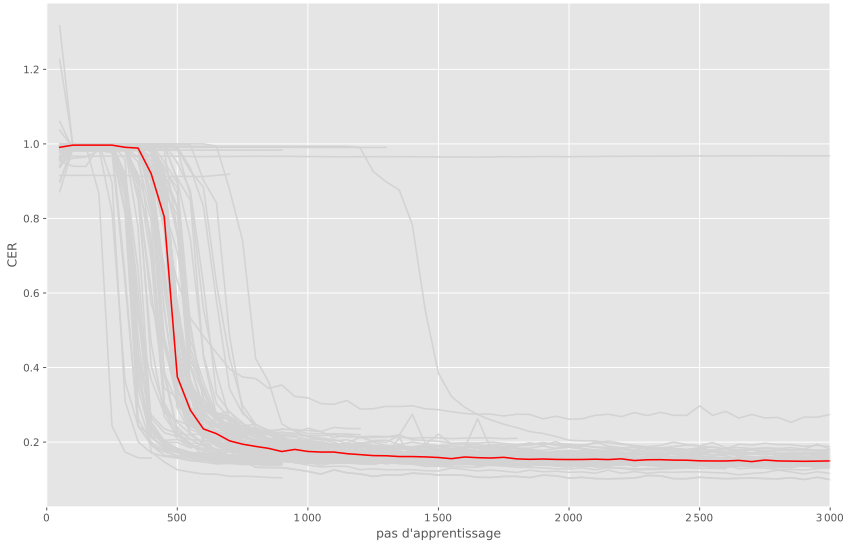

FIGURE 2 – CER sur l'ensemble de validation au cours de différentes optimisations. La courbe rouge correspond à la médiane des CER à chaque étape.

Sur les 91 apprentissages représentés en figure 2, les CER obtenus sur l'ensemble de validation varient entre 8,8 % et 28,8 % (moyenne $14,8$, écart-type $2,16$). La plupart des systèmes appris ont des performances nettement inférieures à celles du système décrit dans nos premières expériences : seuls 6 systèmes ont un CER en validation inférieur à 12,0 % et aucun n'atteint les performances du système de la section 3. Même si tous ces taux d'erreurs ne sont pas directement comparables, ces résultats montrent non seulement que les performances sur l'ensemble de validation sont fortement sensibles aux choix des hyper-paramètres (comme on s'y attendait), mais surtout que la valeur optimale de ceux-ci varie selon les corpus et les configurations. Les CER obtenus sur l'ensemble de validation varient entre 8,8 % et 28,8 % ($M = 14,8$, $S = 2,16$).

(Afin de faciliter la reproduction des expériences, le corpus japhug est rendu disponible en tant que jeu de données Huggingface [4], directement utilisable avec les outils décrits ici.)

Toutefois, comme le montrent les résultats du tableau 2, si l'on applique les différents modèles ob-

3. Notons que les résultats reportés à la figure 2 correspondent aux CER observés durant l'apprentissage sur l'ensemble de validation et non aux résultats du système sur un ensemble de test comme à la figure 1. Ces deux figures ne sont donc pas directement comparables.

4. `https://huggingface.co/datasets/BenjaminGalliot/pangloss`

	①	②	③
CER validation		8,8 %	13,9 %
WER	5,9 %	19,5 %	21,6 %
CER	4,2 %	9,1 %	6,7 %
⊖ ponctuation	1,9 %	6,8 %	4,5 %
⊖ pinyin	0,7 %	2,9 %	4,0 %

TABLEAU 2 – Évaluation détaillée de différents systèmes de transcription phonémique : ① est le système décrit à la section 3, ② le système avec le plus petit CER sur l'ensemble de validation et ③ celui avec le plus petit CER sur l'ensemble de test. Ces deux derniers systèmes prédisent la ponctuation et le symbole @.

tenus au texte corrigé de la section 3.2, la qualité des transcriptions est suffisante pour ne nécessiter qu'un petit nombre de corrections. Ce résultat est d'autant plus remarquable que ces systèmes n'ont été appris que sur 3h de données annotées, une quantité de données « raisonnable » pour l'aide à la documentation linguistique. Il apparaît surtout que les performances des modèles sur l'ensemble de validation ne semblent pas être un indicateur de leur qualité en pratique. Cela complique singulièrement leur sélection et plus généralement leur développement. De manière plus qualitative, nous avons reporté dans le tableau 1 un extrait de la transcription de ce texte par le système décrit à la section 3 et par un système prédisant la ponctuation. Il apparaît que, si le premier système est capable de réaliser une transcription parfaite à l'exception des mots en chinois (romanisés en *pinyin*) et des signes de ponctuation, le second système présente des propriétés qui peuvent être tout à fait intéressantes pour des chaînes de traitement innovantes pour la documentation computationnelle des langues. Tout d'abord, il pose sans erreur les frontières des énoncés (matérialisées par le point), découpage fondamental dans la structure des documents linguistiques tels qu'ils sont encodés dans le format des archives qui les reçoivent. En outre, il reconnaît remarquablement les emprunts chinois, ouvrant la voie à leur identification automatique, qui comporte elle aussi d'importants enjeux pour les études linguistiques comme pour l'application d'outils de traitement automatique des langues.

5 Conclusion

Nous avons décrit dans ce travail la manière dont une approche de type spécialisation (*fine-tuning*) d'un modèle multilingue pouvait être utilisée pour apprendre un système de transcription phonémique automatique pour une langue rare et ainsi réduire l'effort d'annotation des linguistes de terrain. Malgré la grande variabilité des scores obtenus sur un ensemble de validation, nous avons réussi à développer des systèmes dont les prédictions ne nécessitaient qu'un petit nombre de corrections, bien plus faible que celui estimé par le CER.

Ce travail montre l'intérêt de ce type d'approche et ouvre de nombreuses perspectives. En particulier, l'approche nous paraît appeler une extension des expériences à d'autres langues rares (par exemple issue de la collection Pangloss), pour évaluer plus largement son intérêt pour la documentation linguistique. Nous souhaitons également, dans nos travaux futurs, améliorer la qualité des prédictions au niveau des mots, par exemple en intégrant un modèle de langue.

Remerciements

Vifs remerciements aux deux évaluatrices/évaluateurs de ce travail.

Ce travail a bénéficié du soutien financier de l'Agence Nationale de la Recherche (projet « La documentation computationnelle des langues à l'horizon 2025 » [ANR-19-CE38-0015-04] et Labex « Fondements empiriques de la linguistique » [ANR-10-LABX-0083]) et de l'Institut des Langues Rares (ILARA-EPHE).

Une partie importante des ressources linguistiques utilisées dans le présent travail a été collectée dans le cadre du projet « Corpus parallèles en langues himalayennes » [ANR-12-CORP-0006].

References

Adams O., Cohn T., Neubig G., Cruz H., Bird S. & Michaud A. (2018). Evaluation phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan: European Language Resources Association (ELRA).

Aikhenvald A. (2020). Language contact and endangered languages. *The Oxford handbook of language contact*, p. 241–260.

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

Conneau A., Baevski A., Collobert R., Mohamed A. & Auli M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *Computing Research Repository (CoRR)*. https://arxiv.org/abs/2006.13979.

Galliot B., Wisniewski G., Guillaume S., Besacier L., Jacques G., Michaud A., Rossato S., Nguyên M.-C. & Fily M. (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. https://halshs.archives-ouvertes.fr/halshs-03475436.

Godard P., Zanon Boito M., Ondel L., Berard A., Yvon F., Villavicencio A. & Besacier L. (2018). Unsupervised word segmentation from speech with attention. In *Interspeech 2018*, Hyderabad, India.

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

Jacques G. (2021). *A grammar of Japhug*. Number 1 in Comprehensive Grammar Library. Berlin: Language Science Press. https://langsci-press.org/catalog/book/295.

Macaire C. (2021). *Recognizing lexical units in low-resource language contexts with supervised and unsupervised neural networks*. Research report, LACITO (UMR 7107). https://hal.archives-ouvertes.fr/hal-03429051.

Macaire C., Wisniewski G., Guillaume S., Galliot B., Jacques G., Michaud A., Rossato S., Nguyễn M.-C. & Fily M. (2021). Spécialisation de modèles neuronaux pour la transcription phonémique : premiers pas vers la reconnaissance de mots pour les langues rares. In *Journées scientifiques du Groupement de recherche "Linguistique informatique, formelle et de terrain" (GDR LIFT)*, Grenoble. https://halshs.archives-ouvertes.fr/halshs-03475443.

Michaud A., Adams O., Cohn T., Neubig G. & Guillaume S. (2018). Integrating automatic transcription into the language documentation workflow: experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, **12**, 393–429. http://hdl.handle.net/10125/24793.

Michaud A., Adams O., Cox C., Guillaume S., Wisniewski G. & Galliot B. (2020). La transcription du linguiste au miroir de l'intelligence artificielle : réflexions à partir de la transcription phonémique automatique. *Bulletin de la Société de Linguistique de Paris*, **116**(1).

Michaud A., Guillaume S., Jacques G., Mac D.-K., Jacobson M., Pham T.-H. & Deo M. (2016). Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Journées d'Etude de la Parole 2016*, volume 1, p. 155–163. https://halshs.archives-ouvertes.fr/halshs-01341631/.

Moore P. (2018). Re-valuing code-switching: lessons from Kaska narrative performances. In J. Christensen, C. Cox & L. Szabo-Jones, Eds., *Activating the heart: Storytelling, knowledge sharing, and relationship*, Waterloo, Canada: Wilfrid Laurier University Press.

Morris E., Jimerson R. & Prud'hommeaux E. (2021). One size does not fit all in resource-constrained ASR. In *Interspeech 2021*, p. 4354–4358: ISCA.

Muller B., Anastasopoulos A., Sagot B. & Seddah D. (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*, p. 448–462: Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.38.

Newman P. & Ratliff M. (2001). *Linguistic fieldwork*. Cambridge: Cambridge University Press.

Niebuhr O. & Kohler K. J. (2011). Perception of phonetic detail in the identification of highly reduced words. *Journal of Phonetics*, **39**(3), 319–329.

Nitzke, Jeanand Hansen-Schirra S. (2021). *A short guide to post-editing*. Number 16 in Translation and Multilingual Natural Language Processing. Berlin: Language Science Press. https://langsci-press.org/catalog/book/319.

Okabe S., Yvon F. & Besacier L. (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. https://hal.archives-ouvertes.fr/hal-03477475/.

Partanen N., Hämäläinen M. & Klooster T. (2020). Speech recognition for endangered and extinct Samoyedic languages. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*.

Prud'hommeaux E., Jimerson R., Hatcher R. & Michelson K. (2021). Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, **15**, 491–513.

Sperber M., Neubig G., Niehues J., Nakamura S. & Waibel A. (2017). Transcribing against time. *Speech Communication*, **93**, 20–30.

Wisniewski G., Guillaume S. & Michaud A. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In D. Beermann, L. Besacier, S. Sakti & C. Soria, Eds., *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*, p. 306–315, Marseille, France: European Language Resources Association (ELRA). https://halshs.archives-ouvertes.fr/hal-02513914.