Ouvrir aux linguistes “ de terrain ” un accès à la transcription automatique

Guillaume Wisniewski, Alexis Michaud, Benjamin Galliot, Laurent Besacier, Séverine Guillaume, Katya Aplonova, Guillaume Jacques

Ouvrir aux linguistes « de terrain » un accès à la transcription automatique

Guillaume Wisniewski[1], Alexis Michaud[2], Benjamin Galliot[2], Laurent Besacier[3], Séverine Guillaume[2], Katya Aplonova[4] et Guillaume Jacques[5]
(1) Laboratoire de linguistique formelle (LLF), CNRS-Université de Paris, France
(2) Langues et civilisations à tradition orale (LACITO), CNRS-Sorbonne Nouvelle, France
(3) Laboratoire d'informatique de Grenoble (LIG), CNRS-Université Grenoble Alpes, France
(4) Langage, langues et cultures d'Afrique (LLACAN), CNRS-INALCO, France
(5) Centre de recherches linguistiques sur l'Asie orientale (CRLAO), CNRS-EHESS, France
guillaume.wisniewski@u-paris.fr, alexis.michaud@cnrs.fr,
b.g01lyon@gmail.com, laurent.besacier@univ-grenoble-alpes.fr,
severine.guillaume@cnrs.fr, {aplooon|rgyalrongskad}@gmail.com

Résumé

Le traitement automatique de la parole commence à réaliser son fort potentiel pour la documentation des langues en danger. Notre objectif est de mettre à la portée des linguistes « de terrain » des outils de transcription automatique à la pointe des avancées technologiques. Une interface graphique conviviale, Elpis, donne désormais accès à Kaldi et ESPnet, deux bibliothèques de pointe pour le traitement automatique de la parole. Une *recette* ESPnet à utiliser dans Elpis donne d'excellents résultats, aussi bien sur deux jeux de données précédemment utilisés pour entraîner des modèles acoustiques (langues na et chatino) qu'avec deux nouveaux jeux de données (japhug et bashkir). L'interface utilisateur d'Elpis a en outre été dotée de traductions. L'installation est facilitée par *conteneurisation* (en utilisant le logiciel libre Docker), et l'entraînement des modèles est accéléré par l'utilisation de processeurs graphiques (à l'aide de la bibliothèque CUDA).

Abstract

User-friendly automatic transcription of low-resource languages

Natural Language Processing now begins to deliver on its promise for language documentation. This paper reports on progress integrating the speech recognition toolkit ESPnet into Elpis, a web front-end originally designed to provide access to the Kaldi automatic speech recognition toolkit. The goal of this work is to make end-to-end speech recognition models available to language workers via a user-friendly graphical interface. Encouraging results are reported on (i) developing an ESPnet recipe for use in Elpis, with preliminary results on data sets previously used for training acoustic models with the Persephone toolkit along with two new data sets that had not previously been used in speech recognition, and (ii) incorporating ESPnet into Elpis along with user interface enhancements and a CUDA-supported dockerfile.

Mots-clés : documentation linguistique, documentation linguistique assistée par ordinateur, reconnaissance automatique de la parole, science ouverte, linguistique de terrain.

Keywords : Language documentation, Computational Language Documentation, Automatic Speech Recognition, Open Science, linguistic fieldwork.

2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Montrouge, France (Virtuel), 10-11 décembre 2020, pages 82–93,

1 Introduction

La transcription de la parole constitue une dimension importante de la documentation linguistique, en particulier s'agissant de langues et civilisations à tradition orale. Des progrès spectaculaires ont été réalisés en matière de reconnaissance automatique de la parole au cours de la dernière décennie (Hinton et al., 2012 ; Hannun et al., 2014 ; Zeyer et al., 2018 ; Hadian et al., 2018 ; Ravanelli et al., 2019 ; Zhou et al., 2020), y compris de grandes réussites pour des langues peu documentées, pour lesquelles peu de ressources numériques sont disponibles (Besacier et al., 2014 ; Blokland et al., 2015 ; Lim et al., 2018 ; van Esch et al., 2019 ; Hjortnaes et al., 2020). Néanmoins, les technologies de transcription automatique ne sont pas encore exploitées à grande échelle par les linguistes « de terrain » et leurs collaborateurs et collaboratrices. Les logiciels de reconnaissance de la parole sont souvent des prototypes de recherche pour lesquels il n'existe pas d'interface graphique et qui exigent des compétences informatiques peu répandues parmi les utilisateurs potentiels, à commencer par une familiarité avec l'utilisation de la ligne de commande.

L'enjeu pour la documentation linguistique est de taille. En effet, la mise en œuvre, dans le cadre d'archives ouvertes telles que la collection Pangloss (Michailovsky et al., 2014), d'outils de pointe en traitement automatique des langues naturelles apporte une forte impulsion à l'enrichissement des ressources hébergées par ces archives. Le dépôt en archive ouverte, auquel les chercheurs ont accès dans le cadre du dispositif mis en place sous l'égide d'Huma-Num (Jacobson et al., 2015), présente des avantages décisifs en termes de pérennité, mais nombre de chercheurs sont (de façon bien compréhensible) plus sensibles aux enjeux de recherche, à court et moyen terme, qu'aux questions qui concernent la postérité lointaine. L'existence de traitements automatisés pour les corpus déposés dans les archives sonores encouragerait chez les linguistes de terrain un changement d'attitude, dans le sens d'un passage à l'*archivage progressif*. En effet, des archives bien outillées en logiciels de TAL ne prêteraient plus le flanc au soupçon de devenir des « cimetières de données » (*data graveyards :* Gippert et al., 2006, 4, 12-13). Elles deviendraient plutôt des « cliniques de données » (*data clinics*)[1], qui offrent aux déposants un soutien technologique dans l'enrichissement de leurs données. (Un exemple en est fourni par les réalisations présentées aux présentes Journées scientifiques par Cécile Macaire, voir sa communication *Alignement temporel entre transcriptions et audio de données de langue japhug*.) Cela encouragerait les dépôts, parce qu'il y aurait à la clef un fort potentiel d'amélioration des transcriptions et annotations (gloses, traductions). De la sorte, nombre de déposants potentiels seraient portés à franchir le pas et engager une démarche d'archivage, plutôt que remettre à plus tard. Le fait de reporter l'archivage, dans l'attente d'un degré de perfection que les corpus n'atteignent souvent jamais, aboutit en effet à une très forte déperdition de données de langues rares.

Elpis[2] est précisément un outil conçu pour permettre aux linguistes et à leurs collaboratrices et collaborateurs d'avoir accès à un outil de reconnaissance automatique de la parole. Il permet d'entraîner son propre modèle acoustique pour la reconnaissance vocale et de transcrire automatiquement des fichiers audio, au moyen d'une interface graphique (Foley et al., 2018, 2019 ; Adams et al., 2021). Le premier moteur de reconnaissance automatique de la parole « historique » auquel donnait accès Elpis était Kaldi[3] (Povey et al., 2011). Kaldi est une bibliothèque libre, très utilisée dans la communauté de traitement de la parole, qui repose sur une architecture qui hybride modèles de Markov cachés et réseaux de neurones profonds. Elle a permis d'obtenir des résultats à la pointe du progrès dans

1. L'expression est d'un relecteur anonyme pour le colloque Comput-EL4 : *Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

2. `https://github.com/CoEDL/elpis`

3. `https://github.com/kaldi-asr/kaldi`

nombre de tâches de reconnaissance de la parole. Dans ce travail, nous décrivons le processus par lequel une autre bibliothèque de reconnaissance de la parole, plus récente, ESPnet[4] (Watanabe et al., 2018), a été ajoutée à Elpis, aux côtés de Kaldi. La raison pour intégrer ESPnet à Elpis est qu'il s'agit d'un outil largement utilisé, qui rassemble une communauté grandissante et fait l'objet d'un développement logiciel soutenu. Mais surtout, ESPnet implémente une architecture de reconnaissance de la parole plus récente reposant uniquement sur des réseaux de neurones : c'est une approche pouvant être qualifiée de bout en bout (*end-to-end*). Le choix entre plusieurs moteurs *(back-ends)* peut permettre d'obtenir de meilleures performances selon la nature du jeu de données. En effet, nombre de locuteurs, nature des documents, rapport signal/bruit, débit de parole... diffèrent considérablement d'un corpus de langue rare à l'autre, de sorte qu'une grande flexibilité des outils est souhaitable.

Nous commencerons par décrire les changements apportés au logiciel Elpis pour y intégrer le moteur ESPnet et le développement d'une recette ESPnet adaptée aux corpus de langues rares. Une recette est un ensemble de scripts et de fichiers de configuration qui permet de faciliter les différentes étapes de la reconnaissance automatique (préparation des données, extraction de traits, entraînement du modèle, etc.), regroupés dans un script enrobeur (*wrapper*) dans lequel il est aisé de spécifier l'architecture du modèle et ses différents hyperparamètres. Nous décrirons ensuite la mise en œuvre de cette recette pour quatre langues. Enfin, nous discuterons des perspectives pour l'avenir de ce projet.

2 Aperçu des outils et des résultats

2.1 Développements réalisés

L'interface d'Elpis permet de charger des fichiers : voir la Fig. 1. Cette interface a été dotée d'une version française ; il est prévu d'ajouter d'autres langues, en collaboration avec des linguistes de terrain qui ont collecté et transcrit des corpus. Au fil de l'entraînement d'un modèle acoustique (Fig. 2), puis de la transcription de nouveaux fichiers audio (Fig. 3), l'interface montre le journal *(log)*. Même si le détail des messages n'est pas intelligible aux utilisateurs, ils peuvent du moins s'assurer que le processus (qui peut durer jusqu'à plusieurs jours) suit son cours et ne s'est pas bloqué. Si l'outil est déployé par une équipe qui possède des compétences en apprentissage automatique, le journal peut être utilisé pour affiner les paramètres en vue d'améliorer les résultats.

Elpis affiche en outre la liste des mots qui apparaissent dans les fichiers utilisés comme corpus d'entraînement, et indique le nombre d'occurrences : voir Fig. 4.

Une image Docker a en outre été réalisée afin de faciliter l'utilisation du logiciel. Cette image est un conteneur isolé comprenant le logiciel ainsi que toutes ses dépendances (bibliothèques, autres programmes, données, etc.), elle peut être téléchargée ou construite localement (selon une suite d'étapes automatisées). Cette image permet notamment de tirer parti des processeurs graphiques (GPU) afin d'accélérer l'entraînement des modèles grâce à l'utilisation d'une bibliothèque de calcul adaptée (CUDA).

La recette ESPnet, développée spécifiquement pour des corpus de petite taille, est librement disponible en ligne, de même que l'ensemble des outils et données[5]. Les données na et japhug, hébergées

4. https://github.com/espnet/espnet

5. https://github.com/persephone-tools/espnet/commit/1c529eab738cc8e68617aebbae520f7c9c919081

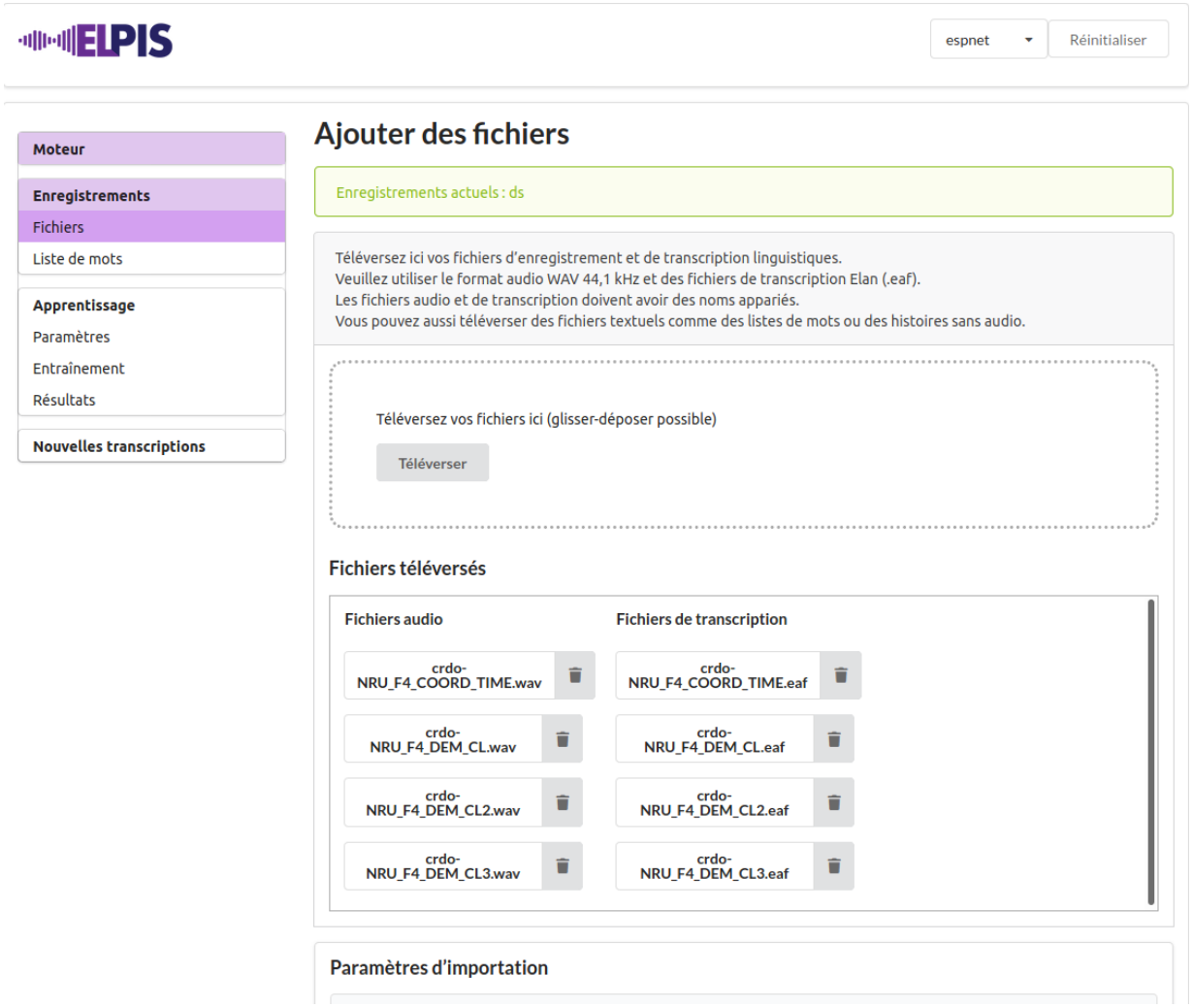

FIG. 1 : Interface d'Elpis. On notera, en haut à droite, le choix de moteur : Kaldi ou ESPnet.

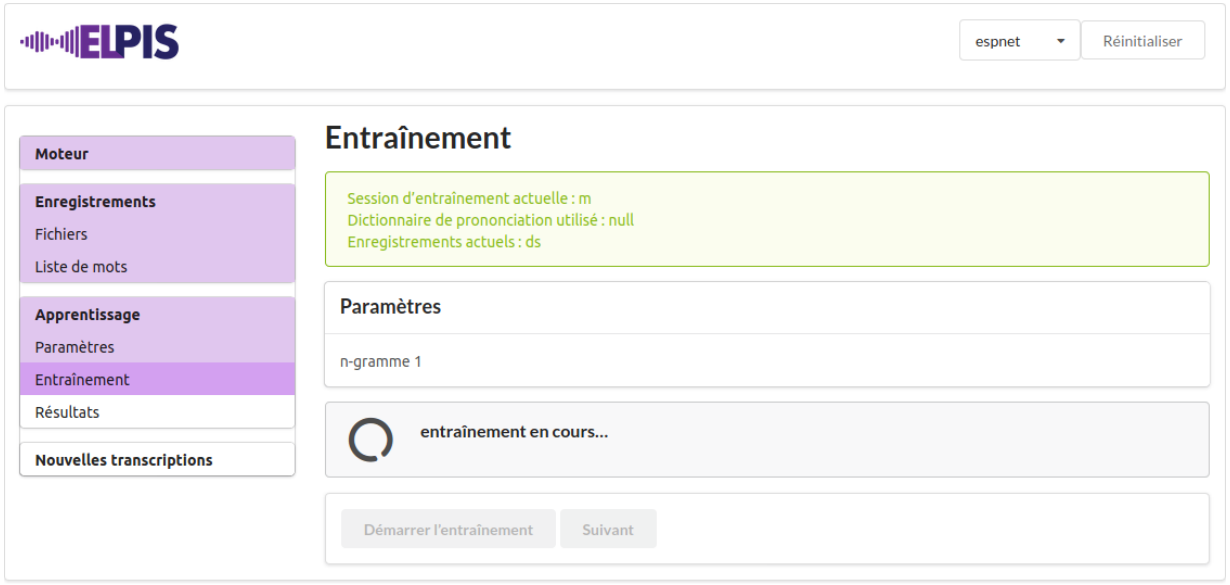

FIG. 2 : Affichage pendant que l'entraînement est en cours.

dans la collection Pangloss, sont converties au format ELAN (seul format d'entrée considéré par Elpis pour le moment) à l'aide d'un script XSLT, baptisé Pangloss-Elpis[6].

6. `https://gitlab.com/lacito/pangloss-elpis`

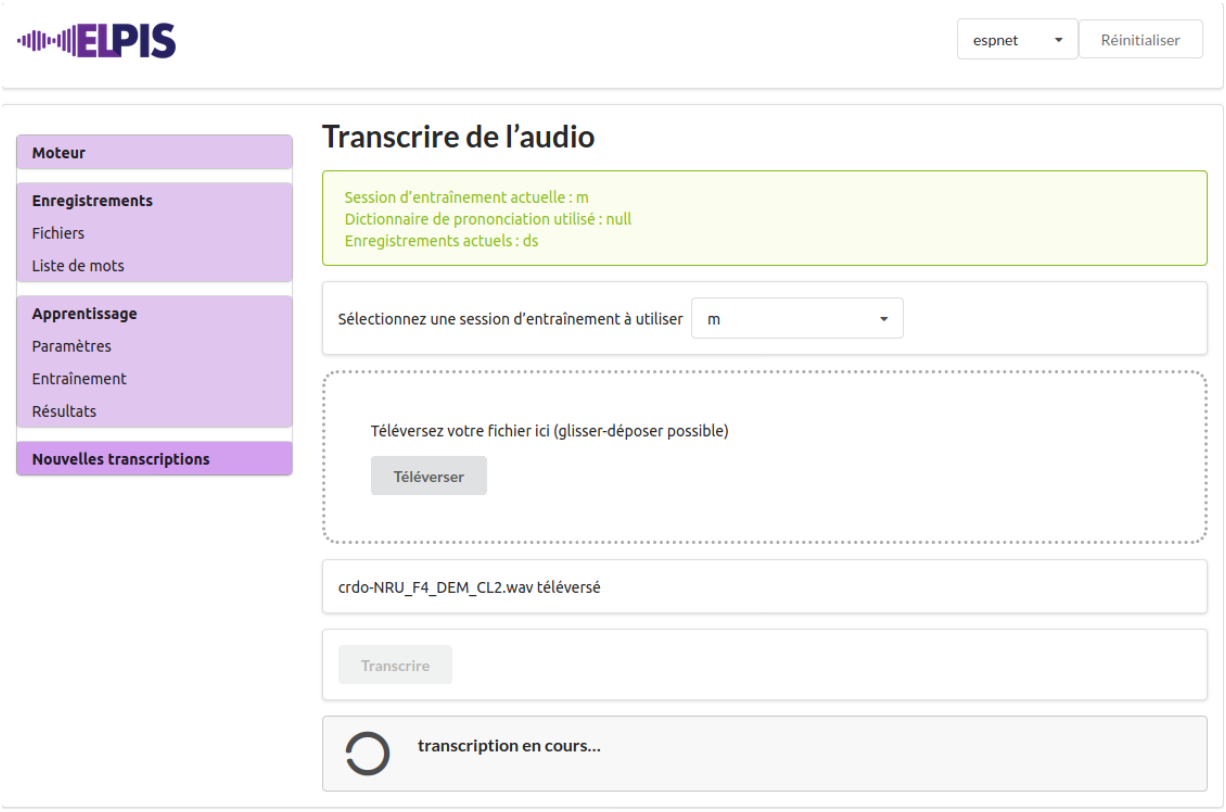

FIG. 3 : Transcription de nouveaux documents audio.

2.2 Résultats

Un bilan des résultats pour quatre langues (cinq jeux de données au total) est présenté dans le tableau 1 ; *taille* désigne la taille du corpus d'entraînement. La figure 5 montre, pour la langue japhug, la baisse du taux d'erreur dans l'identification des phonèmes – ou plus précisément des caractères – selon la taille du corpus d'entraînement, jusqu'à 170 minutes. Des tests sont en cours pour déterminer dans quelle mesure des améliorations sont possibles en utilisant un corpus d'entraînement plus étendu.

Le travail sur le corpus bashkir en est à ses toutes premières étapes et le résultat n'est pas encore probant. Le nombre élevé de locuteurs (36, contre un seul pour les autres corpus : na, chatino et japhug) y est certainement pour beaucoup : le passage du mono-locuteur au multi-locuteurs est une difficulté bien connue pour les systèmes de reconnaissance automatique de la parole, au vu des importantes différences inter-individuelles dans la prononciation. Néanmoins, dans l'ensemble, les résultats obtenus sont encourageants et montrent qu'il est envisageable, dès aujourd'hui, d'utiliser des méthodes de reconnaissance automatique pour faciliter le travail des linguistes de terrain.

Il faut toutefois noter que la qualité des prédictions varie fortement en fonction des langues. Il reste à déterminer dans quelle mesure ces variations tiennent aux propriétés intrinsèques à la langue (taille de l'inventaire phonémique, complexité phonotactique, phénomènes de réduction phonétique liés à la morphosyntaxe, à la structure de l'information...) et dans quelle mesure elles sont liées à des propriétés du corpus (qualité d'enregistrement du signal audio, genres de documents recueillis, nombre de locuteurs). Ce travail prolongera les tests réalisés à ce jour (Adams et al., 2017, 2018 ; Michaud et al., 2018, 2019), qui permettent déjà certaines généralisations (Wisniewski et al., 2020 ; Michaud et al., 2020b,a). De même, des expériences supplémentaires seront nécessaires pour déterminer si une même recette permet d'obtenir des résultats optimaux pour toutes les langues ou s'il sera néces-

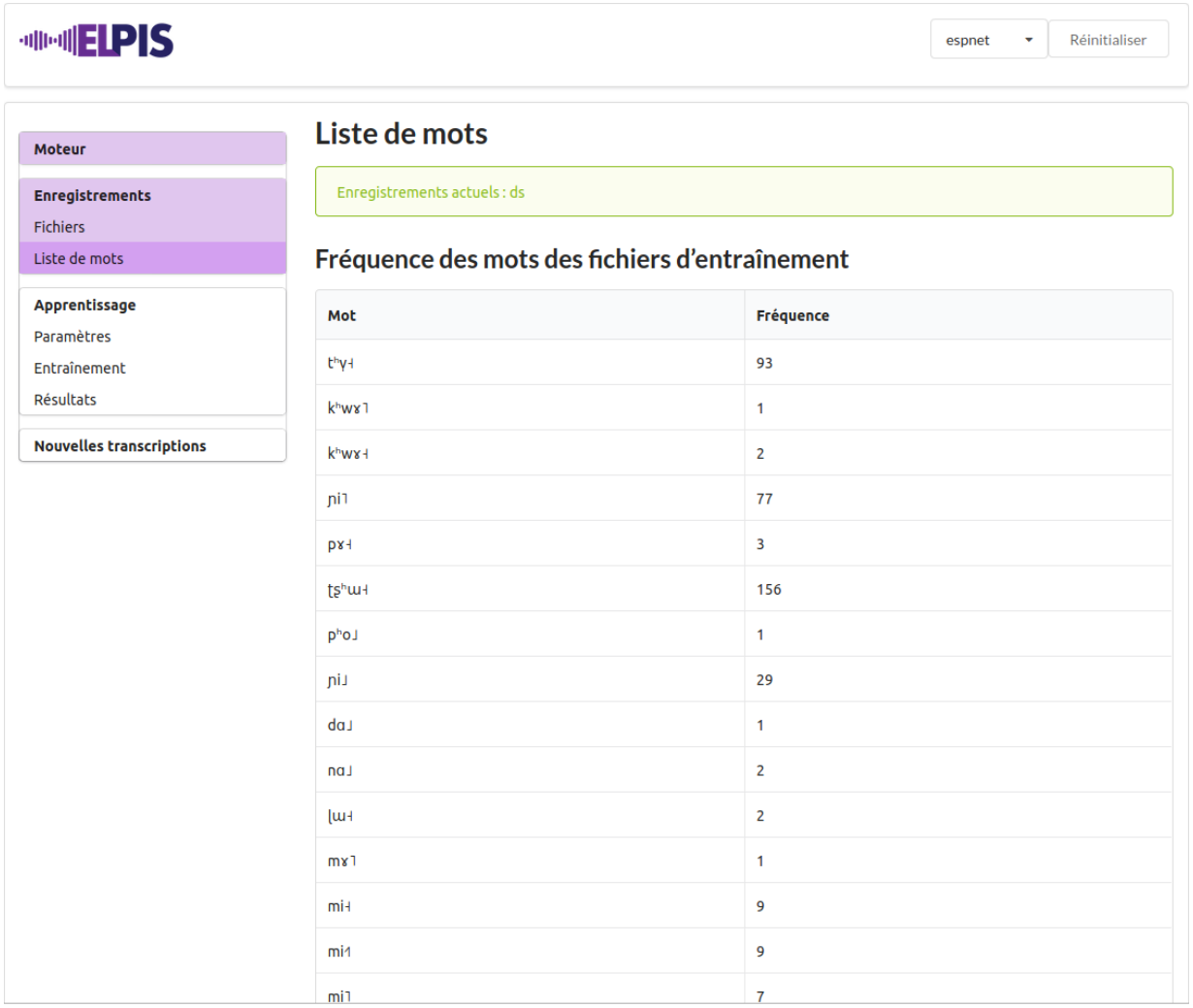

FIG. 4 : Liste des mots du corpus d'entraînement avec mention de leur fréquence d'occurrence.

saire d'adapter l'architecture du réseau de neurones et les hyperparamètres aux spécificités de chaque langue.

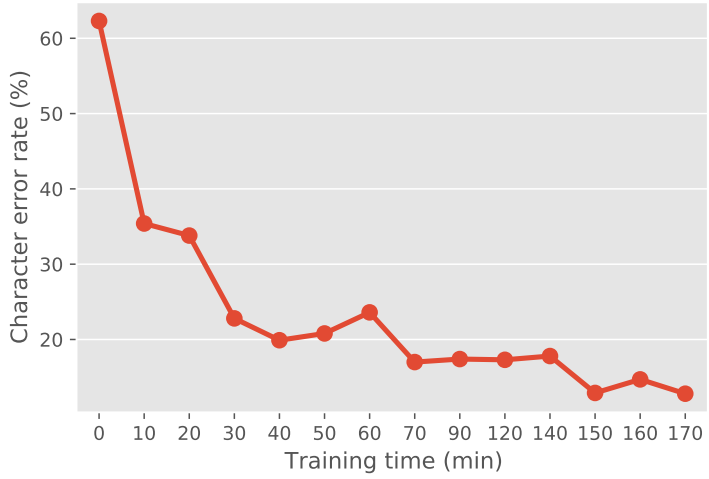

FIG. 5 : Taux d'erreur des graphèmes *(Character Error Rate)* pour le corpus japhug en fonction de la taille du corpus d'entraînement. Moteur : ESPnet.

Langue	Nb locuteurs	Type	Taille (mn)	CER (%)
Na	1	*Récits spontanés*	273	14.5
Na	1	*Expressions élicitées*	188	4.7
Chatino	1	*Parole lue*	81	23.5
Japhug	1	*Récits spontanés*	170	12.8
Bashkir	36	*Récits spontanés*	273	33

TAB. 1 : Résultats de la recette ESPnet sur quatre langues. Les performances sont évaluées par le taux d'erreur des graphèmes *(Character Error Rate)*.

3 Discussion : place d'outils de reconnaissance automatique de la parole dans la documentation linguistique

Un cadre pour une discussion générale au sujet du rôle que peuvent jouer des outils de reconnaissance automatique de la parole dans la documentation linguistique nous est opportunément fourni par l'un des locuteurs invités de la précédente édition des Journées scientifiques du Groupement de recherche LIFT (Orléans, 2019). Dans son exposé, dont une forme développée est sous presse dans la revue *Computational Linguistics* à l'heure où nous écrivons (automne 2020), Steven Bird relève les problèmes que pourrait poser une approche dans laquelle on s'imposerait de généraliser, dans la documentation des langues en voie de disparition, des chaînes de traitement d'outils désormais classiques en reconnaissance de la parole, mais qui seraient mal adaptées aux conditions réelles sur le terrain. La critique de Steven Bird s'adresse notamment à l'emploi d'un modèle acoustique. Passer par l'étape d'une transcription exhaustive en phonèmes (ou autre représentation *segmentale*), au sein de laquelle il faut ensuite identifier les mots, ce serait se créer une difficulté supplémentaire : dans la parole, on ne retrouve pas les phonèmes dans leur réalisation canonique, qui apparaîtraient en une succession sagement linéaire. L'oral est au contraire caractérisé par la variation, les hésitations et reprises, et l'inventivité constante. Steven Bird relève également la diversité des environnements de travail concernés, et des compétences disponibles au sein des groupes qui travaillent à des tâches de documentation linguistique et de sauvegarde des langues. Il préconise une méthode moins linéaire, qu'il dénomme « transcription clairsemée » (« *sparse transcription* ») : annoter ce qui peut l'être aisément, en tirant le meilleur parti des diverses sources d'information disponibles (notamment des traductions), sans s'astreindre d'emblée à une transcription exhaustive et linéaire (Bird, 2020).

Ces réflexions offrent l'occasion de préciser la place qu'occupe un outil tel qu'Elpis dans le paysage des méthodes computationnelles pour la documentation linguistique. Personne ne conteste qu'il serait bien utile de disposer d'une transcription phonémique automatique. En revanche, la question de son intégration dans les chaînes de traitement mérite à l'évidence d'être posée. Steven Bird s'oppose spécifiquement à l'idée selon laquelle la transcription phonémique constituerait une bonne base de départ pour la transcription et l'annotation.

Assurément, la transcription phonémique n'est pas une fin en soi et l'équipe du projet Elpis ne se satisfait pas de parvenir à une transcription en phonèmes : là n'est pas le produit final. L'objectif est évidemment de reconnaître mots et phrases : tel est l'enjeu de la reconnaissance automatique de la parole. Mais il importe de mesurer que transcription phonémique et reconnaissance automatique de la parole ne s'opposent pas : dans les systèmes de reconnaissance purement statistiques de bout en

bout (« end-to-end artificial neural networks »), dont fait partie ESPnet, la distinction n'existe pas au plan technique. C'est là une différence importante, qui distingue ESPnet d'un outil tel que Kaldi. Les méthodes de transcription automatique que nous explorons ne sont pas cantonnées à la reconnaissance d'unités du niveau phonémique. Le fonctionnement du logiciel ESPnet est tout à fait compatible avec un auto-apprentissage *(self-supervised training)* sur des données audio non transcrites, affiné par un ajustement à la langue-cible sur un ensemble de données transcrites (corpus de linguistes tels que ceux utilisés dans notre travail). Or il ne paraît pas absurde, au vu des résultats les plus récents, de penser qu'un outil comme wav2vec 2.0 puisse atteindre des performances remarquables en reconnaissance de la parole (reconnaissance de *mots*) par un processus d'ajustement qui ne repose que sur quelques dizaines de minutes d'enregistrements transcrits de référence, ce qui constitue un ordre de grandeur que savent atteindre les linguistes de terrain. Certes, beaucoup d'expériences ont lieu sur l'anglais, mais il ne nous paraît pas y avoir lieu de douter du potentiel d'extension des mêmes méthodes aux langues rares qui constituent notre objet (voir en particulier Conneau et al., 2020)[7].

En outre, lorsqu'on pèse les avantages respectifs de diverses chaînes de traitement, il est de bonne méthode de tenir compte non seulement des performances actuelles des outils logiciels, mais également de leur potentiel de perfectionnement à court et moyen terme. Avant de conclure qu'il faut se détourner de la chaîne de traitement classique en reconnaissance automatique de la parole (*entraînement sur corpus de référence puis application à la transcription exhaustive de données audio*), il paraît avisé de prendre la mesure des souplesses qu'elle autorise en pratique. Au vu des récents succès de systèmes de reconnaissance de la parole pour des langues en danger (voir en particulier Partanen et al., 2020), il ne paraît pas y avoir lieu de craindre que les chaînes de traitement de la reconnaissance automatique de la parole enferment les linguistes dans une impasse méthodologique.

Clairement, Elpis n'est pas destiné à une utilisation lors des premières étapes de documentation d'une langue. Le choix d'une annotation « clairsemée » (« *sparse transcription* ») paraît particulièrement pertinent dans la situation dans laquelle on se trouve lorsque la langue à décrire n'a pas encore fait l'objet d'une analyse linguistique selon les méthodes de la linguistique de terrain, telles que décrites dans des travaux déjà anciens qui conservent leur actualité (Bouquiaux and Thomas, 1971). Des outils de reconnaissance tels que ceux auxquels Elpis donne accès peuvent en revanche être d'une grande utilité lorsqu'on dispose d'un corpus ciselé par un·e linguiste qui a atteint un certain degré de certitude dans l'analyse de la langue (processus d'analyse qui est toujours en devenir).

En outre, nous faisons l'hypothèse selon laquelle le seuil de taille de corpus à partir duquel une langue peut bénéficier d'outils de traitement automatique de grande qualité va continuer à s'abaisser, de sorte que le cercle des langues pour lesquelles des outils de reconnaissance automatique de la parole puissent être déployé va s'élargir rapidement. Nous pensons que ces progrès pourront être mis à profit quelles que soient les chaînes de traitement adoptées. À mesure du déploiement d'Elpis, nous aurons à cœur de veiller à ce que les utilisateurs ne se trouvent pas enfermés dans le carcan d'une méthode unique, mais que leurs réflexions au fil des expériences contribue au contraire à orienter la manière dont les outils s'adaptent pour répondre aux besoins.

7. L'équipe qui développe l'outil ESPnet étudie actuellement les modalités d'implémentation de ces avancées : voir, par exemple, `https://github.com/espnet/espnet/issues/2609`.

4 Conclusion et perspectives

Les perspectives de déploiement de l'outil paraissent prometteuses. Les projets en cours concernent l'amélioration de la recette et l'amélioration de l'interface. Par ailleurs, des tests sont en cours sur un nouveau jeu de données. En outre, une perspective qui paraît hautement souhaitable serait de proposer Elpis sous forme de service web, sur le modèle de WebMAUS[8]. Les outils développés par l'archive bavaroise de la parole, *Bavarian Speech Archive* (Kisler et al., 2017), ont traité plus de dix millions de fichiers multimédias depuis leur passage en production en 2012, ce qui constitue un modèle de réussite de déploiement à grande échelle d'outils de traitement automatique de la parole.

Remerciements

Le travail décrit ici est réalisé en collaboration par une équipe internationale de chercheurs qui se sont rassemblés au fil du temps autour du projet commun de faciliter les tâches de documentation et description des langues en voie de disparition. Le format du présent document ne permettait pas de faire justice à la liste réelle des contributeurs. Il s'agit au premier chef d'Oliver Adams (qui a identifié ESPnet comme un outil prometteur, et a réalisé son intégration dans Elpis), et de l'équipe du projet Elpis : Nicholas Lambourne, Ben Foley, Rahasya Sanders-Dwyer, Janet Wiles. Nous remercions également Christopher Cox, qui a réalisé un greffon *(plugin)* de transcription automatique pour le logiciel Elan ; Nick Evans, Nick Thieberger, Steven Morey, qui jouent un rôle important dans la coordination internationale du projet Elpis et sa diffusion auprès des linguistes ; et Hilaria Cruz, Martine Adda, Graham Neubig et Nathan Hill pour leur appui. Faute d'espace pour faire figurer tous les collègues qui mériteraient, à divers titres, d'apparaître parmi les auteurs, nous avons fait le choix de ne retenir comme auteurs du présent exposé en français que les participants au groupement de recherche LIFT, organisateur de l'événement. Choix quelque peu arbitraire, mais qui a à tout le moins recueilli le consentement des autres collaborateurs.

Un grand merci aux collègues et amis consultants des langues concernées par les expériences rapportées ici : pour la langue na, il s'agit en particulier de Mme Latami Dashilame et son fils Latami Dashi ; pour la langue japhug, de Tshendzin (Chen Zhen) ; pour la langue tsuut'ina, remerciements particuliers au Bureau du Commissaire à la langue tsuut'ina.

Nous remercions l'Institut des langues rares (ILARA) de l'École Pratique des Hautes Études, l'Université du Queensland et l'*Australian Research Council Centre of Excellence for the Dynamics of Language* pour le soutien financier apporté au développement d'outils de transcription automatique pour la documentation linguistique. Le présent travail est en outre une contribution au projet Labex « Fondements empiriques de la linguistique » (ANR-10-LABX-0083) ainsi qu'au projet « La documentation computationnelle des langues à l'horizon 2025 » (ANR-19-CE38-0015-04).

Références

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC*

8. https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

2018 (Language Resources and Evaluation Conference), pages 3356–3365, Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Adams, O., Cohn, T., Neubig, G., and Michaud, A. (2017). Phonemic transcription of low-resource tonal languages. In *Proceedings of the 2017 Australasian Language Technology Association Workshop (ALTA 2017)*, pages 53–60, Brisbane, Australia. https://halshs.archives-ouvertes.fr/halshs-01656683.

Adams, O., Galliot, B., Wisniewski, G., Lambourne, N., Foley, B., Sanders-Dwyer, R., Wiles, J., Michaud, A., Guillaume, S., Besacier, L., Cox, C., Aplonova, K., Jacques, G., and Hill, N. (2021). User-friendly automatic transcription of low-resource languages: plugging ESPnet into Elpis. In *Proceedings of ComputEL-4: Fourth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Hawai'i. https://halshs.archives-ouvertes.fr/halshs-03030529.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Bird, S. (2020). Sparse transcription. *Computational Linguistics*, pages 1–50.

Blokland, R., Fedina, M., Gerstenberger, C., Partanen, N., Rießler, M., and Wilbur, J. (2015). Language documentation meets language technology. In *Proceedings of the First International Workshop on Computational Linguistics for Uralic Languages - Septentrio Conference Series*, pages 8–18. http://septentrio.uit.no/index.php/SCS/article/view/3457/3386.

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

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

Foley, B., Arnold, J., Coto-Solano, R., Durantin, G., and Ellison, T. M. (2018). Building speech recognition systems for language documentation: the CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, pages 200–204, Gurugram, India. ISCA. https://www.isca-speech.org/archive/SLTU_2018/pdfs/Ben.pdf.

Foley, B., Rakhi, A., Lambourne, N., Buckeridge, N., and Wiles, J. (2019). Elpis, an accessible speech-to-text tool. In *Proceedings of Interspeech 2019*, pages 306–310, Graz. https://www.isca-speech.org/archive/Interspeech_2019/pdfs/8006.pdf.

Gippert, J., Himmelmann, N., and Mosel, U. (2006). Language documentation: What is it and what is it good for. In *Essentials of language documentation*, volume 178, pages 1–30. Walter de Gruyter, Berlin.

Hadian, H., Sameti, H., Povey, D., and Khudanpur, S. (2018). End-to-end speech recognition using lattice-free MMI. In *Interspeech*, pages 12–16. https://danielpovey.com/files/2018_interspeech_end2end.pdf.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et al. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*. https://arxiv.org/abs/1412.5567.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Others (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *Signal Processing Magazine, IEEE*, 29(6):82–97.

Hjortnaes, N., Partanen, N., Rießler, M., and Tyers, F. M. (2020). Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. In *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, pages 31–37, Wien. Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.iwclul-1.5/.

Jacobson, M., Larrousse, N., and Massol, M. (2015). La question de l'archivage des données de la recherche en SHS (Sciences Humaines et Sociales). In *Archives et données de la recherche (ICA/SUV 2014)*, Paris. http://halshs.archives-ouvertes.fr/halshs-01025106.

Kisler, T., Reichel, U., and Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347. ISBN: 0885-2308 Publisher: Elsevier.

Lim, K., Partanen, N., and Poibeau, T. (2018). Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian. In *Proceedings of LREC (International Conference on Language Resources and Evaluation)*, Miyazaki. https://hal.archives-ouvertes.fr/hal-01856178.

Michailovsky, B., Mazaudon, M., Michaud, A., Guillaume, S., François, A., and Adamou, E. (2014). Documenting and researching endangered languages: the Pangloss Collection. *Language Documentation and Conservation*, 8:119–135. https://halshs.archives-ouvertes.fr/halshs-01003734.

Michaud, A., Adams, O., Cohn, T., Neubig, G., and 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., and Guillaume, S. (2019). Phonetic lessons from automatic phonemic transcription: preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. In *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*, Melbourne. https://halshs.archives-ouvertes.fr/halshs-02059313.

Michaud, A., Adams, O., Cox, C., Guillaume, S., Wisniewski, G., and Galliot, B. (2020a). 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). https://halshs.archives-ouvertes.fr/halshs-02881731/.

Michaud, A., Adams, O., Guillaume, S., and Wisniewski, G. (2020b). Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo). In Benzitoun, C., Braud,

C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *Actes de la 6e conférence conjointe Journées d'Études sur la Parole, Traitement Automatique des Langues Naturelles, Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, pages 451–462, Nancy, France. ATALA. https://hal.archives-ouvertes.fr/hal-02798572.

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

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. https://infoscience.epfl.ch/record/192584/files/Povey_ASRU2011_2011.pdf.

Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6465–6469. IEEE. https://arxiv.org/abs/1811.07453.

van Esch, D., Foley, B., and San, N. (2019). Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1, Honolulu, Hawai'i. https://computel-workshop.org/wp-content/uploads/2019/02/CEL3_book_papers_draft.pdf#page=26.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., and Chen, N. (2018). ESPnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*. https://arxiv.org/abs/1804.00015.

Wisniewski, G., Guillaume, S., and Michaud, A. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In Beermann, D., Besacier, L., Sakti, S., and Soria, C., editors, *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). https://halshs.archives-ouvertes.fr/hal-02513914.

Zeyer, A., Irie, K., Schlüter, R., and Ney, H. (2018). Improved training of end-to-end attention models for speech recognition. https://arxiv.org/abs/1805.03294.

Zhou, W., Michel, W., Irie, K., Kitza, M., Schlüter, R., and Ney, H. (2020). The RWTH ASR system for TED-LIUM Release 2: Improving Hybrid HMM with SpecAugment. https://arxiv.org/abs/2004.00960.