La transcription du linguiste au miroir de l'intelligence artificielle : réflexions à partir de la transcription phonémique automatique

Alexis Michaud, Oliver Adams, Christopher Cox, Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot

Accepté pour publication dans le *Bulletin de la Société de Linguistique de Paris*.
Auteurs : Alexis Michaud, Oliver Adams, Christopher Cox, Séverine Guillaume, Guillaume Wisniewski, Benjamin Galliot.

LA TRANSCRIPTION DU LINGUISTE AU MIROIR DE L'INTELLIGENCE ARTIFICIELLE : RÉFLEXIONS À PARTIR DE LA TRANSCRIPTION PHONÉMIQUE AUTOMATIQUE

RÉSUMÉ. - Les systèmes de reconnaissance automatique de la parole atteignent désormais des degrés de précision élevés sur la base d'un corpus d'entraînement limité à deux ou trois heures d'enregistrements transcrits (pour un système mono-locuteur), au lieu de dizaines d'heures pour les outils antérieurs. Au-delà de l'intérêt pratique que présentent ces avancées technologiques pour les tâches de documentation linguistique, se pose la question de leur apport pour la réflexion du linguiste. En effet, le logiciel réalise son entraînement sur la base de transcriptions fournies en entrée par le linguiste, transcriptions qui reposent sur un ensemble d'hypothèses plus ou moins élaborées, et plus ou moins explicites. Le modèle acoustique, décalqué (par des méthodes statistiques) de l'écrit du linguiste, peut-il être interrogé par ce dernier, en un jeu de miroir ? Que peut nous apprendre la confrontation ainsi renouvelée avec le signal acoustique ? La présente étude s'appuie sur des exemples de langue na (famille sino-tibétaine) pour illustrer la façon dont l'analyse d'erreurs permet une confrontation renouvelée avec les données. Quelques réflexions au sujet d'expériences de transcription automatique de la langue tsuut'ina (famille dene) sont également présentées.

1. Introduction

1.1. Linguistique et Traitement Automatique des Langues

Le Traitement Automatique des Langues a connu d'importants progrès au cours des deux dernières décennies, mais les collaborations entre informaticiens et linguistes ont été moins intenses qu'on ne pourrait le souhaiter. Les gains de performance ont été principalement

obtenus en tirant parti d'une puissance de calcul sans cesse accrue, ainsi que de nouveaux outils statistiques. Une modélisation par règles, comme en propose le linguiste (Vaissière 1971), donnerait de moins bons résultats que des outils statistiques qui ne supposent aucune connaissance explicite de leur objet. C'est ce à quoi fait allusion le mot de Frederick Jelinek selon lequel « à chaque fois qu'un linguiste quitte le groupe de travail, le taux de réussite du logiciel de reconnaissance de la parole augmente », qui remonterait à 1988 (Martin & Jurafsky 2009: 83), et qui continue de faire partie de la tradition orale du domaine de la reconnaissance automatique de la parole. Dans cette perspective, la reconnaissance automatique de la parole en mode non supervisé (« end-to-end ») apparaît comme un beau défi technologique à relever : le Grand Jeu du traitement automatique des langues contemporain, en quelque sorte. Moins il est nécessaire que l'humain intervienne, plus grande est la prouesse informatique. En lisant le compte-rendu des efforts pour entraîner un modèle de reconnaissance automatique en ne recourant *qu'à une supervision limitée, voire sans supervision aucune* (« with limited or no supervision » : Kahn et al. 2019), on devine bien le désir qui anime les chercheurs : atteindre un mode non supervisé, *via* des étapes où le degré de supervision décroît progressivement. Ainsi, dans le domaine de la phonétique-phonologie, des modèles pré-entraînés sur des enregistrements audio (signaux de parole) dépourvus d'annotation permettent de parvenir à des performances améliorées (Schneider et al. 2019) : la statistique permet de déceler certaines régularités dans les transitions entre états acoustiques successifs, qui facilite la reconnaissance automatique. Au-delà de l'anglais, langue sur laquelle se concentre une grande partie de l'attention des groupes de recherche en Traitement Automatique des Langues, les méthodes de pré-apprentissage dites *non supervisées* donnent également des résultats concluants pour d'autres langues, au point d'être présentées comme prometteuses pour les langues les plus diverses, y compris celles pour lesquelles on dispose de ressources linguistiques numérisées en nombre assez limité (Rivière et al. 2020).

Mais si l'on y regarde de près, il ressort que l'apprentissage non supervisé est employé comme complément à un apprentissage qui demeure supervisé : « we apply unsupervised pre-training to improve supervised speech recognition » (Schneider et al. 2019). Le pré-apprentissage non supervisé ne permet pas de faire l'économie d'une seconde étape qui consiste en un apprentissage sur données annotées. Comme on pourrait s'y attendre, l'apprentissage statistique supervisé donne de meilleurs résultats que l'apprentissage non supervisé (Wu et al., 2018 ; voir aussi Jimerson & Prud'hommeaux, 2018).

Dans ce contexte, des collaborations renouvelées entre linguistes et spécialistes du Traitement Automatique des Langues Naturelles sont clairement porteuses d'enjeux importants. Le dialogue interdisciplinaire demeure aussi pertinent que jamais à l'ère de l'apprentissage machine.

Les réflexions présentées ici s'inscrivent dans le cadre d'un travail d'équipe qui associe informaticiens et linguistes afin de concevoir et déployer des outils innovants, et d'en tirer parti pour la recherche.

1.2. La transcription automatique pour langues « peu dotées informatiquement »

L'utilisation d'outils de transcription automatique constitue un enjeu considérable pour la documentation linguistique, dans un contexte d'urgence : il s'agit d'accélérer le travail de collecte et de description d'une diversité linguistique mondiale en déclin rapide (Littell et al. 2018; Thieberger 2017; van Esch, Foley & San 2019). En outre, doter une langue (et la culture qu'elle véhicule) d'outils numériques est important pour qu'elle conserve une certaine place dans un contexte où elle se trouve minoritaire et marginalisée (Soria, Besacier & Pretorius 2018). Les recherches exploratoires menées par un jeune chercheur en informatique, Oliver Adams (Adams 2017), ont fourni l'occasion de nouer une collaboration autour d'un outil de transcription automatique, `Persephone` (Michaud et al. 2018)[1], et de constater que les systèmes de reconnaissance automatique de la parole atteignent désormais des degrés de précision élevés sur la base d'un corpus d'entraînement limité à deux ou trois heures d'enregistrements transcrits (pour un système mono-locuteur), au lieu de dizaines d'heures pour les outils antérieurs. A titre d'exemple, les figures 1 et 2 fournissent les résultats (inédits à ce jour) obtenus en mai 2020 lors d'une première application de l'outil `ESPnet` (Watanabe et al. 2018) à deux jeux de données : la figure 1 représente les résultats obtenus sur un jeu de données de langue na (présentée au §2.1) et la figure 2 les résultats pour des données de langue chatino. (Au sujet du corpus chatino, voir Ćavar, Cavar & Cruz 2016; au sujet du système tonal chatino, particulièrement complexe, voir Cruz 2011; Cruz & Woodbury 2014.) Le taux d'erreur est de l'ordre de 10% pour le na et 18% pour le chatino. La différence de taille du jeu de données (50 minutes pour le chatino, contre 220 minutes pour le na) explique dans une large mesure cette différence dans les résultats. Ces résultats (moins de 20% d'erreurs) sont d'autant plus remarquables

[1] On signalera également des exposés en vidéo (en anglais) au sujet de l'outil de transcription `Persephone` (https://www.youtube.com/watch?v=IwWKqxQ7Qng) et de son intégration dans le logiciel de documentation linguistique `ELAN` (https://www.youtube.com/watch?v=-pDOEqRpZKs).

qu'ils sont obtenus en première passe, avant les ajustements spécifiques au jeu de données traité (d'une part le prétraitement des données afin d'éliminer quelques scories – telles que des incohérences de la notation –, d'autre part le réglage des « hyperparamètres » de l'outil logiciel), lesquels permettent d'espérer des progrès supplémentaires dans la qualité des résultats.

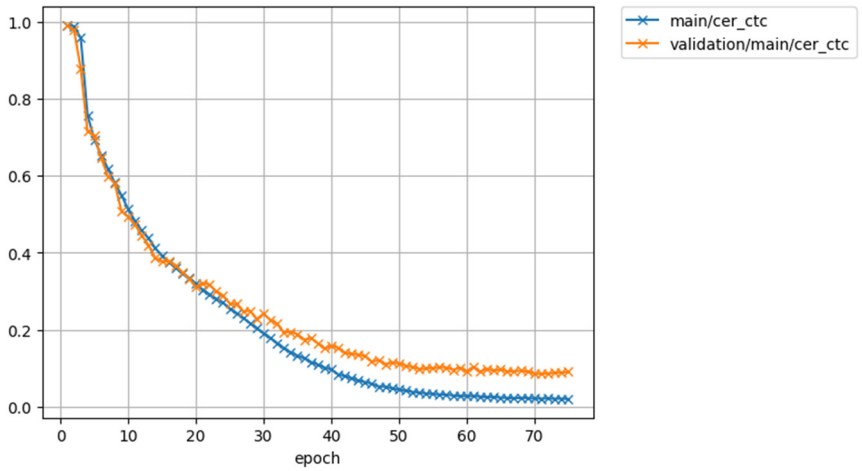

FIGURE 1 : taux d'erreur de reconnaissance des phonèmes et tons de langue na de Yongning avec le logiciel `ESPnet`. En abscisse : nombre d'itérations (en anglais *epochs*) du processus d'entraînement du modèle. Ligne foncée : résultats sur données d'entraînement. Ligne claire : résultats sur données de test.

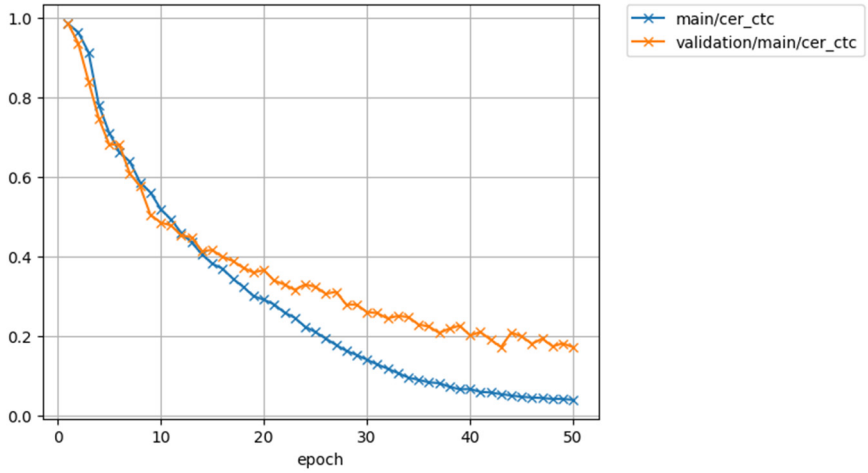

FIGURE 2 : taux d'erreur de reconnaissance des phonèmes et tons de langue chatino avec le logiciel `ESPnet`. En abscisse : nombre d'itérations (en anglais *epochs*) du processus d'entraînement du modèle. Ligne foncée : résultats sur données d'entraînement. Ligne claire : résultats sur données de test.

Pour dire quelques mots au sujet de l'architecture informatique sur laquelle repose les outils : les logiciels de reconnaissance automatique de la parole dont il est ici question appartiennent au vaste ensemble des algorithmes qui recourent à la fonction objective dite de *classification temporelle connectionniste*, CTC (Graves et al., 2013 ; en français, on consultera notamment Tomashenko & Estève, 2018). Le signal audio est soumis à une décomposition fréquentielle par banc de filtres (ce qui revient, pour l'essentiel, à ce que livre une représentation spectrographique), par fenêtres de 10 ms (avec chevauchement). Les traits ainsi extraits sont fournis en entrée à un réseau multi-couche de neurones artificiels récurrents. Une caractéristique importante de cette approche est que le modèle ne contient pas d'hypothèses concernant l'alignement temporel des unités reconnues, de sorte que « l'alignement entre les éléments d'entrée et les étiquettes de sortie est inconnu » (Tomashenko & Estève 2018: 561). Cette propriété du modèle permet de traiter, outre les phonèmes, l'information non segmentale, telle que les tons lexicaux ainsi que tout autre type d'événement figuré dans la transcription fournie en entrée, par exemple un découpage en mots prosodiques.

Au plan du développement logiciel, un travail en cours vise à proposer une interface utilisateur unique qui permette au linguiste « de terrain » d'appliquer à ses données (corpus d'entraînement composé de fichiers audio transcrits, et corpus d'application composé de fichiers audio non transcrits) toute une gamme d'outils : `Persephone` (Adams et al. 2017; Adams et al. 2018), `wav2letter++` (Pratap et al. 2018), `ESPnet` (Watanabe et al. 2018), `KALDI` (Povey et al. 2011)... En effet, au vu de l'ampleur des différences que présentent entre elles les langues naturelles au plan phonético-phonologique comme à d'autres niveaux (morphosyntaxe, structure de l'information...), il paraît vraisemblable que des outils logiciels différents soient plus ou moins performants selon la langue et le type de corpus. Certains donneront de meilleurs résultats que d'autres pour le traitement des tons ou de l'accent, pour des données transcrites dans une notation qui s'éloigne plus ou moins du niveau phonémique, pour le passage d'un mode mono-locuteur à un mode multi-locuteurs, et ainsi de suite. Ces questions empiriques font à l'heure actuelle l'objet d'expérimentations (voir notamment Wisniewski, Guillaume & Michaud 2020). Le développement d'un outil utilisable par des linguistes « de terrain » se fait actuellement en collaboration avec les auteurs de l'outils Elpis (Foley et al. 2018; Foley et al. 2019). À moyen terme, l'objectif est de parvenir, au-delà de la transcription des phonèmes, à un système complet de reconnaissance automatique de la

parole, qui repose sur un modèle de langage (à l'exemple de Hjortnaes et al. 2020).

1.3. Apports des outils informatiques à une réflexion épistémologique

Dans le présent travail, qui fait suite à deux exposés à des colloques (Michaud et al. 2019; Michaud et al. 2020), l'accent n'est pas mis sur l'utilité pratique des outils de Traitement Automatique des Langues, mais sur les possibilités qu'offre la transcription automatique pour la recherche linguistique. En 1951, Alan Turing formulait l'intuition selon laquelle « les tentatives de création de machines pensantes nous seront d'une grande aide pour découvrir comment nous pensons nous-mêmes »[2]. Soixante-dix ans plus tard, ce domaine a acquis une grande importance dans la recherche en informatique : mois après mois, de nouvelles techniques sont proposées pour déterminer ce que « savent » les modèles statistiques (Hohman et al. 2019; Jiang et al. 2019; Lapuschkin et al. 2019; Montavon, Samek & Müller 2017). Dans l'interprétation des résultats, il faut bien sûr savoir raison garder (Gomez-Marin 2017), mais sans pour autant se priver de suivre les chercheurs en informatique dans leurs explorations en rapide renouvellement.

Spécifiquement, la présente communication se veut une réflexion au sujet des enseignements qui se peuvent tirer de l'utilisation d'outils de transcription phonémique automatique. Le recours à une transcription phonémique automatique fournit l'occasion d'une confrontation renouvelée avec le signal acoustique, dans la mesure où le logiciel n'a pas accès à des connaissances de plus haut niveau (reconnaissance des mots, compréhension de l'intention de communication). Cette configuration n'est pas sans rappeler la posture adoptée par le phonéticien-phonologue, qui choisit de suspendre le flux ordinaire de la communication orale pour prêter attention à la forme phonique de la parole. Selon le mot de Lewis Carroll, le locuteur n'a à se soucier que du sens, et les sons suivront d'eux-mêmes : « take care of the sense, and the sounds will take care of themselves » (Carroll 1866: 133, détournant le précepte de frugalité « take care of the pence, and the pounds will take care of themselves »). En revanche, la phonétique expérimentale recourt à des outils qui séparent des dimensions que l'oreille humaine perçoit de façon solidaire, telles que fréquence fondamentale, durée, intensité, et pente spectrale (Gendrot 2003). C'est donc sans aucun paradoxe qu'on peut défendre le point de vue selon lequel les réflexions

[2] Texte original : « I believe that the attempt to make a thinking machine will help us greatly in finding out how we think ourselves ». Voir : http://www.turingarchive.org/browse.php/B/5, consulté le 2 février 2020.

au sujet du fonctionnement des outils de reconnaissance automatiques des sons de la parole peuvent être pertinentes pour la phonétique-phonologie, à laquelle elles ouvrent de nouvelles portes (voir par exemple Shi et al. 2015).

2. Méthode

2.1. La langue na de Yongning

Le na, aussi appelé mosuo (appellation officielle en Chine depuis 1990) ou narua (appellation retenue dans le registre *Ethnologue* depuis 2010), est une langue sino-tibétaine parlée dans la plaine de Yongning et autour du lac Lugu, à la frontière entre les provinces chinoises du Yunnan et du Sichuan (Lidz 2010). L'inventaire phonémique et la phonotactique du dialecte de Yongning ont été établis au fil d'enquêtes de terrain de 2006 à 2016 (Michaud 2008; Michaud 2017: 447–486). Une caractéristique saillante de la langue na est le rôle de premier plan qu'y jouent les tons : rôle morpho-phonologique aussi bien que lexical (Michaud 2017). Une documentation a été réunie (Michaud et al. 2012), qui comporte en particulier vingt-sept récits transcrits et traduits en français.

2.2. La validation croisée : comparaison des transcriptions générées automatiquement avec les transcriptions manuelles

La méthode employée afin de comparer les transcriptions générées automatiquement avec les transcriptions manuelles est la *validation croisée*. L'un des documents est retranché du corpus disponible pour une langue donnée (documents, essentiellement des récits, transcrits manuellement par le linguiste), et un modèle acoustique est entraîné sur le reste du corpus puis appliqué sur le texte qui avait été réservé à cet effet. Cette procédure est appliquée successivement à chacun des vingt-sept documents du corpus. Des fichiers (au format PDF) ont été générés en mettant en valeur, pour chaque phrase (unité <S> du format de la Collection Pangloss : voir Michailovsky et al. 2014), les écarts entre la transcription de référence (manuelle) et la transcription générée automatiquement. Pour la langue na, les documents du corpus d'entraînement peuvent être consultés dans la Collection Pangloss (Michaud et al. 2016)[3], et les documents qui présentent la comparaison des transcriptions générées automatiquement avec les transcriptions manuelles peuvent être consultés dans un dépôt GitHub[4] organisé de façon chronologique (2017 et 2018 : résultats obtenus au moyen du

[3] https://pangloss.cnrs.fr/corpus/list_rsc.php?lg=Na&name=na
[4] https://github.com/alexis-michaud/na/tree/master/AutomaticTranscription

logiciel Persephone ; 2020 : résultats obtenus au moyen du logiciel ESPnet).

2.3. Choix d'une analyse qualitative

L'évaluation d'un modèle acoustique s'effectue généralement en quantifiant le taux d'erreur par comparaison avec une transcription de référence produite (ou du moins vérifiée) par un annotateur humain. Dans le travail décrit ici, une évaluation globale a été réalisée, qui conclut, pour le logiciel Persephone, à des taux d'erreur de l'ordre de 17% pour la langue na (Adams et al. 2018), et pour ESPnet, de l'ordre de 10%, comme mentionné en introduction (voir figure 1 ci-dessus), en net progrès par rapport à une étude-pilote réalisée sur les mêmes données au moyen de CMU-Sphinx (Do, Michaud & Castelli 2014). Au-delà de ce résultat général encourageant, nous avons choisi d'examiner des exemples détaillés, rangés en quelques catégories simples, plutôt que d'aborder l'analyse d'erreurs au moyen d'outils statistiques.

3. Résultats concernant la langue na de Yongning

Le tableau 1 présente une matrice de confusion pour un modèle acoustique entraîné avec le logiciel ESPnet. La ligne 'diff' indique la différence entre le nombre d'occurrences dans la transcription de référence (celle établie à la main par le linguiste) et dans la transcription automatique, ce qui fournit une indication concernant une tendance globale à « entendre » telle ou telle unité trop souvent, ou trop peu souvent. Le tableau est relativement équilibré : il n'y a pas de biais flagrant par lequel telle consonne ou voyelle apparaîtrait beaucoup plus souvent qu'une autre dans les transcriptions automatiques. À titre indicatif, les confusions constatées dans plus de 10 cas sont signalées par une cellule grisée. Le tableau 1 montre que les erreurs de l'outil statistique (ESPnet) se concentrent essentiellement sur les voyelles, et particulièrement les voyelles hautes. On constate aussi des confusions entre l'occlusive dentale /t/ et la post-alvéolaire/rétroflexe /ʈ/. Bref, des confusions qui paraissent clairement en rapport avec des réalités acoustiques : le faible degré de différence phonétique entre voyelles hautes (souvent dévoisées en na, ce qui ne facilite pas leur bonne reconnaissance) et entre des consonnes de point d'articulation proche.

	i	ɯ	o	t	v	m	d	ɹ	e	ɛ	æ	z	s	k	l	z	ə	t	h	n	total
i	1396	18	14	0	1	0	0	7	12	0	1	2	0	0	1	0	0	0	0	1	1453
ɯ	22	948	22	0	2	0	0	9	6	0	1	0	0	0	0	0	1	1	0	0	1012
o	11	14	927	0	3	0	0	6	1	0	3	0	0	0	0	0	0	0	0	0	965
t	0	0	0	747	0	0	8	0	0	6	0	0	0	0	0	0	0	8	0	1	770
v	4	0	2	0	627	0	1	0	0	2	0	3	1	0	0	0	0	0	0	0	640
m	0	0	1	0	1	494	0	0	0	0	0	0	0	0	0	0	7	0	0	2	505
d	0	1	0	2	1	0	466	0	0	1	1	0	0	0	0	0	0	0	0	0	472
ɹ	8	8	19	5	1	0	0	376	4	0	5	0	0	0	0	0	1	0	0	0	427
e	12	7	3	2	0	1	0	2	371	0	1	0	0	0	0	0	1	0	0	0	400
ɛ	0	0	0	2	0	0	0	0	0	371	0	0	0	1	0	0	0	0	0	0	374
æ	2	5	1	0	0	0	0	4	0	1	343	0	0	0	0	0	2	0	0	0	358
z	0	0	0	0	0	0	0	0	0	0	0	362	0	0	3	6	0	0	0	0	371
s	0	0	0	0	0	0	0	0	0	1	0	0	357	0	1	1	0	0	0	0	360
k	0	0	0	2	0	0	0	0	0	1	0	0	0	351	0	0	0	0	1	0	355
l	2	0	0	0	1	2	0	0	0	0	1	3	0	0	304	1	0	0	0	1	315
z	1	1	0	0	3	1	0	0	0	0	0	7	3	0	2	295	0	0	0	1	314
ə	2	0	0	0	1	6	0	0	0	0	2	0	0	1	0	1	246	0	0	0	259
t	0	0	0	13	0	0	0	0	0	0	0	0	0	0	0	0	0	280	0	0	293
h	0	0	0	0	4	0	0	0	0	2	0	0	0	0	1	0	0	0	266	0	273
n	1	0	0	0	0	2	0	0	1	0	0	0	0	0	2	0	0	0	0	247	253
total	1461	1002	989	773	645	506	475	404	395	385	358	377	361	353	314	304	258	289	267	253	
diff	8	-10	24	3	5	1	3	-23	-5	11	0	6	1	-2	-1	-10	-1	-4	-6	0	

TABLEAU 1 : Matrice de confusion pour un échantillon de test de transcription phonémique automatique avec le logiciel `ESPnet`. Chaque ligne correspond à un phonème. Les colonnes indiquent le nombre de cas dans lesquels le phonème en question a été catégorisé d'une certaine façon. La ligne 'diff' indique la différence entre le nombre d'occurrences dans la transcription de référence (celle établie à la main par le linguiste) et dans la transcription automatique.

Le tableau 1 fournit en outre matière à réflexion au sujet des affriquées. En na, les combinaisons de symboles /tɕ/ et /dʑ/ notent en effet des consonnes affriquées. Les deux composantes de ces digraphes (*t* et *ɕ*, *d* et *ʑ*) ne sont pas à proprement parler des phonèmes : elles ne possèdent pas d'autonomie au plan structurel, et ne sont, en ce sens, pas identiques aux monographes. Parmi les deux outils logiciels de transcription automatique dont il est ici question, `Persephone` et `ESPnet`, le premier effectuait un prétraitement des données qui comportait une segmentation en phonèmes, par lequel la séquence /tɕi/, par exemple, se trouvait segmentée en /tɕ/ + /i/ ; le second ne comportait pas de segmentation de ce type, de sorte que les deux parties du digraphe /tɕ/ n'étaient pas traités de façon distincte des phonèmes monographes /t/ et /ɕ/. Les /t/ et /ɕ/, /d/ et /ʑ/, /ʈ/ et /ʂ/ du tableau 1 représentent donc à la fois les phonèmes correspondants, et les occurrences de ces symboles au sein de digraphes. Pour autant, les taux d'erreur restent faibles. Ici se pose une question d'interprétation. Ces unités sont bien reconnues par un logiciel auquel n'était pas fournie en entrée l'information linguistique selon laquelle certains phonèmes sont notés par des digraphes et d'autres par des monographes. Faut-il en conclure que la partie fricative du phonème /tɕ/ présenterait une proximité phonétique particulièrement élevée avec la fricative /ɕ/ ?

Cette conclusion ne va en fait nullement de soi. Il est utile de revenir ici sur le mode de fonctionnement de l'outil logiciel. Celui-ci prend en compte une fenêtre relativement étendue du signal acoustique (de l'ordre d'une dizaine de secondes), qui s'étend donc bien au-delà de la durée du son considéré (par exemple un bruit de friction). Le caractère contextuel de l'outil signifie que la différence phonologique entre les trois attaques syllabiques que constituent l'affriquée /tɕ/, l'occlusive /t/ et la fricative /ɕ/ est accessible au logiciel *en tant que régularité statistique*. Les états acoustiques associés au symbole /ɕ/ n'ont pas besoin d'être invariants, du moment qu'un élément du contexte (un autre symbole dans la chaîne, tel qu'un /t/ qui précède) permet de faire la part des choses. Autrement dit, une grande diversité acoustique ne pose pas de problème à l'outil si elle est fortement corrélée au contexte. Au-delà des résultats qui figurent dans le tableau 1 (concernant, rappelons-le, l'outil `ESPnet`), la comparaison de tests réalisés avec l'outil `Persephone` confirme qu'un prétraitement par lequel est fournie la liste des unités à reconnaître (les phonèmes) n'a qu'une incidence limitée sur les performances de l'outil (Wisniewski, Guillaume & Michaud 2020).

Après ces quelques réflexions générales, venons-en à l'analyse détaillée d'erreurs constatées dans la transcription automatique.

3.1. Harmonie vocalique

La transcription de /tʰɣ˩qo˩/ ‘là-bas’ comme /tʰo˩qo˩/ (Funeral.44)[5] témoigne de la présence d’une tendance phonétique à l’harmonie vocalique, dont on trouve de façon sporadique des traces lexicalisées en na ainsi que dans les autres langues naïques (Michaud 2017: 466–467). Si le phénomène était uniforme, on s’attendrait à ce que l’outil de transcription, qui est sensible au contexte phonologique large, puisse le compenser, de la même façon qu’il compense de lui-même les phénomènes de coarticulation entre phonèmes adjacents. Le fait que des catégorisations erronées sporadiques soient constatées dans les transcriptions générées automatiquement suggère que le degré d’harmonie vocalique (assimilation de timbre d’une voyelle – plus précisément d’une rime – à celle de la syllabe suivante) varie notablement d’un exemple à l’autre. Le modèle automatique est de nature statistique, de sorte qu’il s’acquitte mal de tâches d’identification dans lesquelles la statistique phonétique à elle seule n’est pas un guide suffisant. L’harmonie vocalique en na a partie liée avec les processus de lexicalisation : le figement progressif d’expressions courantes. Un exemple typique en est /boɻhɑ˥/, « pâtée des cochons », qui dans certains dialectes du naxi (langue proche du na) est devenu /bɑɻhɑ˩/, tandis que la voyelle des deux syllabes conserve une différence de timbre dans d’autres dialectes. (Les correspondances tonales entre na et naxi sont particulièrement complexes, et ne seront pas abordées ici.)

Ainsi, les erreurs de transcription automatique peuvent guider le linguiste vers les zones du système sonore de la langue qui reçoivent une influence notable de facteurs non prédictibles d’après la connaissance du phonème et de son environnement phonémique.

3.2. Oppositions phonémiques et syllabiques

Si certaines substitutions ont lieu au niveau du phonème, comme l’illustrent les exemples d’harmonie vocalique évoqués ci-dessus, il n’est pas rare d’observer la substitution d’une syllabe entière. Ainsi de /jiɻ/ ‘champ’ reconnu comme /hĩɻ/ ‘être humain’ (Agriculture.29[6]), de /dɣɻpʰæɻ/ reconnu comme /dʑoɻpʰæɻ/ ou /ɖɯɻpʰæɻ/ (Funeral.44[7]), de /ɣɻ/ reconnu comme /ɭɯɻ/ (BuriedAlive3.51[8]). Ces substitutions à l’échelle de la syllabe entière fournissent confirmation de l’observation

[5] Les références aux textes en ligne sont fournies dans le format suivant : <titre du document>.<numéro de phrase>. Un lien DOI (Digital Object Identifier) est en outre fourni en note, de façon à permettre un accès direct à l’exemple cité, et ainsi une navigation plus fluide entre publications, données et outils (Vasile et al. 2020). Par exemple, pour Funeral.44 : https://doi.org/10.24397/pangloss-0004572#S44

[6] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004441#S29

[7] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004572#S44

[8] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004539#S51

selon laquelle les syllabes du na (majoritairement de structure très simple : consonne+voyelle, et un ton) connaissent une forte coarticulation. La consonne et la voyelle appariées deviennent fortement coarticulées, et leurs caractéristiques ont tendance à se répandre sur l'ensemble de la syllabe. Eugénie Henderson avait observé une propension au déplacement de traits phonologiques au sein de la syllabe dans d'autres langues phonologiquement monosyllabiques d'Asie (Henderson 1985). Dans les langues du groupe naïque, auquel appartient le na, la coarticulation des monosyllabes CV tend à créer des unités compactes qui se plient de moins en moins commodément à une analyse simple en deux phonèmes distincts, jusqu'au point où la syllabe devient monophonémique. Ainsi, 'os' se dit /ʈ̃ɨ/, mot dans lequel divers traits (nasalité, caractère rhotique, rétroflexion...) se concentrent en un seul segment. Ce phénomène atteint également un degré extrême dans les voyelles apicalisées, courantes dans la région (Baron 1974), ainsi que dans les nasales syllabiques (Bradley 1989: 150; Björverud 1998: 8). Les erreurs de transcription automatique montrent que, pour des syllabes dont la division demeure claire, consonne et voyelle entretiennent déjà une forte solidarité. Les syllabes du na de Yongning, fortement érodées (Jacques & Michaud 2011), sont en chemin vers une association toujours plus étroite entre consonne et voyelle.

3.3. Accent d'insistance et réalisation des phonèmes

Il est bien connu que les phénomènes prosodiques ont une influence sur la réalisation des phonèmes. En na, l'accent d'insistance placé sur une syllabe peut amener à une erreur de transcription. Cela peut paraître paradoxal : une syllabe sous accent d'insistance sera hyper-articulée, selon le terme de Lindblom (1990), et sa bonne reconnaissance devrait en être facilitée, en comparaison de la même syllabe articulée avec moins d'énergie. Mais l'outil logiciel employé pour la transcription fonctionne sur des bases statistiques, non sur la base d'une réalisation canonique. C'est ainsi que le mot 'terre', /ʈʂeɬ/, se retrouve transcrit comme /ʈʂʰɯɬ/ dans un contexte où il est porteur d'un net accent d'insistance (Seeds.30[9]). À l'inverse, la consonne aspirée /tsʰ/ est identifiée comme un simple /ts/, sans aspiration, dans un nom propre quadrisyllabique, au sein duquel elle a peu d'espace pour se déployer, de sorte qu'elle est phonétiquement réduite (BuriedAlive3.51[10]).

Cet exemple fournit une transition opportune pour aborder les noms propres quadrisyllabiques.

[9] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004548#S30
[10] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004539#S51

3.4. La situation particulière des noms propres quadrisyllabiques

Un exemple de transcription automatique par le logiciel `Persephone`, suivi de la transcription de référence (manuelle) en vis-à-vis, est fourni dans le tableau 2. Les gloses figurent en exemple (1).

æ̃˩	tʂʰe˦	ɖɯ**˦**	**mæ**˩		tʂʰ**ɯ˦**	**bi˦**	mæ˩	pi˦	dʑo˩	*transcription manuelle*
ɻ̩˩	tʂʰe˦	ɖɯ**˩**	**mɑ**˩	**ɻ̩˩**	tʂʰ**e˦**	**ɖɯ˩**	**mɑ**˩	pi˦	dʑo˩	*transcription automatique*

TABLEAU 2 : Exemple de validation croisée de transcription automatique par le logiciel `Persephone`. Les différences sont mises en valeur en gras.

(1) ɻ̩˩ tʂʰe˦ ɖɯ˩ mɑ˩ pi˦ dʑo˩
Erchei-Ddeema *(nom propre)* dire TOPICALISATEUR
Elle a crié : « Erchei-Ddeema ! Erchei-Ddeema ! »
(Enterrée vive.13[11])

Dans ce court passage, on relève des erreurs de transcription sur les deux occurrences du nom proche Erchei-Ddeema (le nom d'un des principaux protagonistes du récit). La forme phonémique de ce nom est /ɻ̩˩ tʂʰe˦ ɖɯ˩ mɑ˩/. Au vu du taux d'erreur globalement faible (de l'ordre de 17%), il est frappant d'observer neuf erreurs sur les tons, les consonnes et les voyelles en l'espace d'à peine huit syllabes. L'examen des onze occurrences de ce nom propre dans le texte (reproduites ci-dessous : tableau 3) révèle qu'aucune n'est exempte d'erreurs.

pæ˩ tʂʰ**ɯ˥** ɖɯ**˦** m**ɣ**˩	**æ̃**˩ tʂʰe˦ ɖɯ**˦** mæ˩	**∅** tʂʰ**ɯ˦** **bi˦** mæ˩
ɑ˩ tʂʰe˦ ɖɯ**˦** m**ɤ**˩	**∅** **tʰi˥** ɖɯ**˦** mɑ˩	**æ˩** tʂʰ**ɯ˦** ɖɯ**˦** m**ɤ˦**
ɻ̩˩ tʂʰe˦ ɖɯ**˦** m**ɤ˦**	ɻ̩**˦** tʂʰ**ɯ˦** **dʑɯ˦** m**ɤ˦**	**æ˩** tʂʰ**ɯ˦** ɖɯ**˦** m**ɤ˦**
ɻ̩˩ tʂʰ**ɯ˦** ɖɯ**˦** m**ɤ˦**	**æ˩** tʂʰ**ɯ˦** ɖɯ**˦** m**ɤ˦**	

TABLEAU 3 : Transcription automatique des onze occurrences du nom propre Erchei-Ddeema /ɻ̩˩ tʂʰe˦ ɖɯ˩ mɑ˩/. Le symbole de l'ensemble vide ∅ indique une syllabe manquante.

La première syllabe, l'approximante syllabique /ɻ̩/, est identifiée comme une voyelle dans six cas, et manque tout à fait dans deux cas.

[11] Accès direct à cet exemple : https://doi.org/10.24397/pangloss-0004537#S13

Ce qui la distingue d'une voyelle (dans les réalisations qu'on dira, selon ses préférences théoriques, *canoniques* ou *hyperarticulées*) est essentiellement la rétroflexion, laquelle se manifeste au plan acoustique par un abaissement du troisième formant, jusqu'à des valeurs de l'ordre de 2 000 Hz, nettement inférieures à toutes les voyelles. Son identification comme une voyelle ouverte suggère que le degré phonétique de rétroflexion / rhotacisation est inférieur, dans ces exemples, à la moyenne statistique.

Le défaut de reconnaissance de ce segment, dans deux cas, tient vraisemblablement à sa coalescence phonétique avec une voyelle qui précède. La structure syllabique (C)V du na de Yongning place en hiatus le noyau de toute syllabe dépourvue de consonne initiale. Un spectrogramme est proposé en figure 3. Il révèle un bref passage glottalisé, qui signale vraisemblablement le découpage en constituants (Dilley & Shattuck-Hufnagel, 1996 ; Kuang, 2017, p. 3218) et qui contribue peut-être à masquer la baisse du troisième formant qui, pour l'œil du phonéticien, signale un mouvement articulatoire que n'explique pas la coarticulation avec la consonne affriquée qui suit.

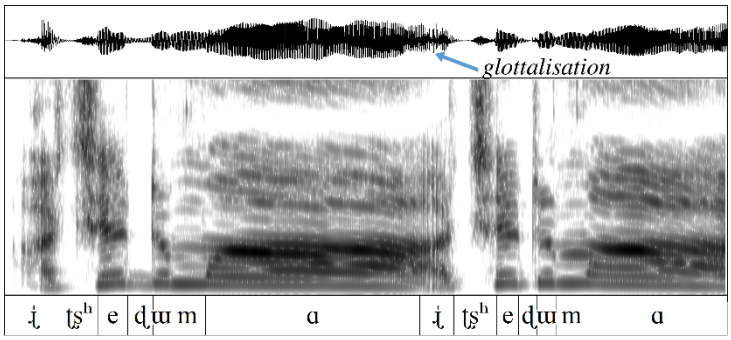

FIGURE 3 : Spectrogramme correspondant à l'exemple (1).

La voyelle de la seconde syllabe du nom /ɻ̩˩ ʈʂʰe˦ ɖɯ˩ mɑ˩/ est identifiée comme un /ɯ/ dans la majorité des cas. En langue na, la voyelle /ɯ/ possède un allophone apical après les fricatives rétroflexes et affriquées : ainsi, /ʈʂʰɯ/ est réalisé [ʈʂʰʐ̩] (ou, si l'on adopte les symboles proposés par Chao Yuen-ren, [ʈʂʰɻ̩]. (Au sujet de ces segments mi-voyelle mi-consonne, voir Shao & Ridouane 2018 et références citées.) L'identification de la voyelle comme /ɯ/ plutôt que /e/ peut donc être interprétée comme la conséquence d'une hypo-articulation de la voyelle. Le mouvement de la langue en direction d'une cible [e] est moins ample que dans la configuration moyenne telle

qu'elle est extraite du corpus d'entraînement par le logiciel de transcription automatique. La langue demeure proche de la configuration adoptée pour la consonne [t͡sʰ].

La troisième des quatre syllabes du nom est, dans ces exemples, moins affectée que les autres, mais son ton est systématiquement identifié comme Moyen (˧) et non Bas (˩). Au plan acoustique, l'examen des données révèle que le schéma tonal /L.M.L.L/ du quadrisyllabe est réalisé avec des valeurs de f_0 plus élevées sur les deuxième et troisième syllabes que sur les première et dernière. Cette observation fait penser aux schémas observés au niveau du mot dans les langues polysyllabiques, et c'est sur cette similarité que nous allons nous appuyer pour proposer une interprétation.

En langue na, les racines lexicales sont monosyllabiques, du fait d'une érosion phonologique très poussée au fil de l'histoire de la langue (Jacques & Michaud 2011). Ces racines monosyllabiques se recombinent en disyllabes par des processus morphologiques de composition et d'affixation, de sorte que les disyllabes sont abondamment attestés dans le lexique, en particulier parmi les noms (Michaud 2012). Les disyllabes fournissent, à leur tour, une base pour la formation de mots plus longs. Les mots de quatre syllabes ou plus représentent environ 6% du lexique enregistré à ce jour (Michaud 2015), et leur fréquence d'occurrence dans les vingt-sept textes transcrits est du même ordre (5,5%). Les quadrisyllabes sont donc marginaux en termes de distribution statistique. Il paraît vraisemblable que le modèle acoustique créé au moyen du logiciel `Persephone` fasse la part belle aux transitions acoustiques telles qu'elles sont réalisées sur les monosyllabes et les disyllabes. Le degré de précision avec lequel est réalisé chacun des phonèmes d'un mot court, donc pauvre en matériau phonologique, a toutes chances d'être plus élevé que pour des mots plus longs.

Il n'y a rien là de bien nouveau : cette tendance était déjà relevée par Marguerite Durand (1930), et les phonéticiens-phonologues qui s'intéressent à la typologie prosodique (tons et accents) ont maintes occasions de l'observer à l'œuvre. L'éclairage qu'apportent les résultats tirés d'expériences de transcription automatique n'en est pas moins intéressant : ces résultats ouvrent de nouvelles perspectives pour l'étude de la hiérarchie entre les multiples facteurs qui entrent en jeu dans les phénomènes de variation allophonique – laquelle recouvre, *in fine*, le domaine entier de la variation intonative (Vaissière 2004).

Il faut souligner ici que différents outils logiciels pour la transcription automatique n'attirent pas nécessairement l'attention sur les mêmes phénomènes. L'outil `ESPnet`, dont les premiers tests pour la langue na sont encore tout récents, présente un taux d'erreur global

plus bas que `Persephone`, et il n'y a pas systématiquement d'erreur sur les mots quadrisyllabiques, de sorte que le constat d'une spécificité des mots longs est moins flagrant. Ainsi, l'une des occurrences du prénom *Erchei Ddeema* (/ɻɭʈʂʰeɬ-ɖɯɹmɑɹ/) dans la transcription automatique par `ESPnet` est exempte d'erreurs[12], même si les deux autres occurrences dans le corpus de test comportent des erreurs d'identification vocalique (/ɻɭʈʂʰeɬ-ɖɯɬmɤɬ/ dans un cas[13], /ɻæɬʈʂʰɯɬ-.../ dans l'autre[14]). Le prénom *Nobbu Ci'er* (/noɹbʏɬ-tsʰɯɹɻɭ/) est bien identifié à chacune de ses quatre occurrences (*modulo* de légères inexactitudes concernant les tons). De la sorte, divers outils logiciels offrent une vision différente des mêmes données.

3.5. Réduction des mots-outils

Fréquence lexicale et nature grammaticale (« mot plein » par opposition à « mot-outil », avec toutes les nuances intermédiaires des charges sémantiques et des degrés de grammaticalisation) constituent des facteurs de variation intonative (allophonique) d'importance variable d'une langue à l'autre. La préposition vietnamienne *cho* 'pour' (DAT), homophone du verbe 'donner' dont elle est issue, en diffère moins au plan phonétique que ne le laisserait attendre l'exemple des langues comme le français ou l'anglais (Brunelle, Chow & Nguyễn 2015). On peut s'attendre à ce que les morphèmes qui connaissent une réduction phonétique soient moins bien reconnus, voire tout simplement omis dans des transcriptions automatiques.

Les observations qualitatives réalisées au sujet des transcriptions automatiques de la langue na suggèrent que la différence entre mots pleins et mots grammaticaux n'est pas particulièrement saillante dans cette langue. Ces observations (qui restent à quantifier) amènent à formuler l'hypothèse selon laquelle la longueur d'un mot a une incidence plus forte sur la façon dont chacun de ses phonèmes est prononcé (dans le corpus considéré) que le statut grammatical du mot (classe morphosyntaxique) et sa fréquence.

On peut espérer que la poursuite de ces observations apporte une contribution de nature typologique aux questions de variation allophonique et de « prosodie articulatoire » (Fougeron 1999; Fougeron 2001).

[12] https://doi.org/10.24397/pangloss-0004537#S129
[13] https://doi.org/10.24397/pangloss-0004537#S149
[14] https://doi.org/10.24397/pangloss-0004539#S21

4. Réflexions au sujet de la transcription automatique appliquée à la langue tsuut'ina

Le tsuut'ina est une langue dene (athabasque) parlée dans le sud de l'Alberta, au Canada. Son vocalisme a été analysé, dans des études linguistiques dont la tradition remonte à près d'un siècle, en termes de quatre voyelles phonémiques, *i, a, o, u* (en phonétique : /ɪ a ɒ ʊ/) (Li 1930 ; Cook 1984). Mais la pertinence synchronique de cette analyse pour la langue tsuut'ina dans son état contemporain a récemment été remise en question par des études acoustiques. L'analyse d'enregistrements de locuteurs tsuut'ina ne fait pas ressortir la différence attendue entre les deux voyelles basses, /a/ et /ɒ/. Une interprétation possible du constat phonétique d'une grande proximité acoustique entre les voyelles /a/ et /ɒ/ est que les quatre voyelles demeureraient néanmoins des phonèmes distincts : telle est la voie choisie par Barreda (2011). En revanche, Sims (2010; 2011) émet l'hypothèse selon laquelle ces différences seraient désormais des réalisations allophoniques d'un unique phonème vocalique bas, lequel fluctuerait entre [ɒ] (dans tous les cas où la voyelle est longue, et dans les cas où la voyelle est brève mais précède une consonne dorsale) et [a] (dans les autres cas). La proximité particulière entre ces deux voyelles dans la langue telle qu'elle est parlée aujourd'hui affleure dans les observations épilinguistiques des locuteurs natifs du tsuut'ina, qui témoignent que la distinction entre le *a* et le *o* de l'orthographe tsuut'ina (les /a/ et /ɒ/ des linguistes) ne va pas de soi. Certains expriment des doutes sur la solidité synchronique d'une opposition qui leur paraît un vestige d'un autre temps.

Ainsi, lorsque des données audio de langue tsuut'ina ont été employées pour l'entraînement d'un modèle acoustique avec le logiciel `Persephone`, rien ne garantissait que l'opposition *a* - *o* y fût présente avec un degré de régularité suffisant pour que la statistique la fasse ressortir. Les transcriptions fournies en entrée pour l'entraînement du modèle étaient orthographiques, et reflétaient donc la distribution lexicale de l'opposition phonologique entre /a/ et /ɒ/. L'emploi d'une transcription orthographique revenait à faire l'hypothèse que l'opposition demeurait pertinente. L'entraînement d'un modèle acoustique sur des données transcrites de façon orthographiques permettait donc de soumettre cette hypothèse à une forme originale de vérification. Si les voyelles /a/ et /ɒ/ (les *a* et *o* de l'orthographe tsuut'ina) ne sont pas distinguées par un locuteur de façon conséquente, un modèle statistique ne sera pas en mesure de les distinguer de manière cohérente.

Or il s'avère que le modèle acoustique entraîné à partir des données tsuut'ina distingue étonnamment bien les deux phonèmes /a/ et /ɒ/, tant dans les contextes où il s'agit de voyelles brèves que dans ceux où ce sont des voyelles longues. Cette observation fournit un argument en faveur de l'hypothèse selon laquelle le locuteur tsuut'ina qui a participé aux enregistrements perpétue bel et bien cette opposition phonémique. Il n'est pas du tout impossible, en principe, qu'un modèle acoustique puisse surpasser le linguiste dans l'exercice qui consiste à déterminer si telle ou telle opposition phonémique conserve sa pertinence.

Une précaution méthodologique s'impose ici. Comme expliqué brièvement au §2, le modèle acoustique tient compte du contexte. Il est donc théoriquement possible que le réseau de neurones artificiel ait appris à transcrire tantôt /a/ et tantôt /ɒ/ en fonction du contexte d'occurrence de ces sons : en s'appuyant sur leurs contextes d'apparition dans les textes fournis en entrée (dans le corpus d'apprentissage), plutôt que sur des caractéristiques acoustiques repérées au sein du signal audio aligné avec ces mêmes textes. Il importe donc de déterminer si les [a] et [ɒ] ne seraient pas de simples allophones prédictibles à partir du contexte phonémique dans les données d'apprentissage. Afin de déterminer si tel est le cas – si la différence est simplement allophonique, et transcrite par l'outil `Persephone` à partir de l'environnement consonantique et de la quantité vocalique –, le cas à examiner de près est celui des voyelles longues, où tout contraste phonétique serait neutralisé (Sims 2010). Or il s'avère que, dans les documents tsuut'ina dont il est question ici, le contexte segmental n'est pas suffisant pour parvenir au résultat observé. Les transcriptions automatiques présentent une classification des voyelles /a/ et /ɒ/ qui est exacte (au sens où la transcription automatique coïncide avec la forme phonologique canonique du mot, consignée dans les dictionnaires) même dans le cas des voyelles longues, donc dans un contexte où /a/ et /ɒ/ apparaissent dans le même environnement segmental. A titre d'exemple, en fin de mot, on observe aussi bien /a/ (par exemple dans *chák'aa* /tʃʰák'aː/ 'côte (partie du corps)') que /ɒ/ (dans *k'oo* /k'ɒː/ 'récent').

Ces exemples autorisent à conclure que le contexte phonologique ne permet pas à lui seul de distinguer les voyelles /a/ et /ɒ/. Il n'en reste pas moins plausible que le modèle créé par `Persephone` fasse quelque usage de la distribution statistique de ces voyelles en fonction du contexte segmental. Dans les cas où les indices acoustiques ne font pas pencher la décision vers l'une de ces deux voyelles à l'exclusion de l'autre, il est vraisemblable que la solution proposée repose sur la fréquence relative de ces voyelles dans le contexte segmental concerné.

Une exploration systématique de cette question constitue l'une des nombreuses pistes possibles pour la poursuite des recherches.

En guise de bilan d'étape au sujet de la transcription automatique de la langue tsuut'ina, on relèvera la double utilité de l'outil, au plan pratique et au plan de la réflexion linguistique. La capacité de l'outil à distinguer des sons acoustiquement proches rend service lors de la production de transcriptions de nouveaux enregistrements. Outre cette utilité pratique pour les tâches de documentation linguistique, la transcription automatique éclaire d'un jour renouvelé un point d'analyse phonémique de la langue, en offrant une source d'information complémentaire des intuitions des locuteurs et des compétences du linguiste.

5. Conclusion et perspectives

Les travaux présentés ici en sont à leurs débuts. L'étape qui consistera à s'élever au-dessus d'observations de détail telles que celles relatées ci-dessus, pour tisser ensemble tous les brins, ne pourra réellement être atteinte qu'au moyen de méthodes plus systématiques (méthodes statistiques, elles aussi) pour la catégorisation des erreurs de transcription. Il nous paraîtrait donc prématuré de prétendre proposer une discussion générale. En revanche, il paraît d'ores et déjà possible de conclure que l'emploi de techniques de Traitement Automatique des Langues Naturelles dans le contexte de la documentation linguistique (« linguistique de terrain ») livre des bénéfices dès les premières étapes de la collaboration entre informaticiens et linguistes. Entre autres perspectives pour la suite du travail, on mentionnera l'extraction d'information à partir des modèles acoustiques générés par apprentissage statistique. L'apprentissage machine suit des procédures qui ne sont pas celles des phonéticiens-phonologues, mais il ne paraît pas impossible de mettre en rapport les probabilités calculées par le modèle avec des variables qui soient interprétables. L'étude des modèles acoustiques pourrait, en particulier, fournir un appui dans l'entreprise qui consiste à caractériser les phonèmes d'une langue en termes de propriétés acoustiques (Vaissière 2011a; Vaissière 2011b) et articulatoires (Stavness et al. 2012), et ainsi parvenir à un degré de précision nettement supérieur à celui que permet l'Alphabet Phonétique International.

6. Remerciements

Un grand merci aux collègues et amis consultants de langue na (en particulier Mme Latami Dashilame et son fils Latami Dashi) et de

langue tsuut'ina (notamment le Bureau du Commissaire à la langue tsuut'ina) pour leur soutien à ce travail. Merci à Martine Adda, Laurent Besacier et Graham Neubig pour leurs conseils et leur soutien.

Nos vifs remerciements au Comité d'organisation de la Journée scientifique de la Société de Linguistique de Paris du 26 janvier 2019 « Corpus, analyses quantitatives et modèles linguistiques » : Annie Rialland, Benoît Sagot et Catherine Schnedecker.

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).

BIBLIOGRAPHIE

Adams, Oliver. 2017. *Automatic understanding of unwritten languages*. Melbourne: The University of Melbourne Ph.D.

Adams, Oliver, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird & Alexis Michaud. 2018. Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. Miyazaki. https://halshs.archives-ouvertes.fr/halshs-01709648.

Adams, Oliver, Trevor Cohn, Graham Neubig & Alexis Michaud. 2017. Phonemic transcription of low-resource tonal languages. In *Proceedings of ALTA 2017 (Australasian Language Technology Association Workshop)*, 53–60. Brisbane. https://halshs.archives-ouvertes.fr/halshs-01656683.

Baron, Stephen P. 1974. On the tip of many tongues: Apical vowels across Sino-Tibetan. In. Georgia State University, Atlanta. https://halshs.archives-ouvertes.fr/halshs-01400987.

Barreda, Santiago. 2011. The Tsuut'ina vocalic system. *Rochester Working Papers in the Language Sciences* 6. 1–10.

Björverud, Susanna. 1998. *A Grammar of Lalo*. Lund: Lund University.

Bradley, David. 1989. Nasals and nasality in Loloish. In David Bradley, Eugénie J.A. Henderson & Martine Mazaudon (eds.), *Prosodic Analysis and Asian Linguistics: to honour R.K. Sprigg*, 143–155. Canberra: Pacific Linguistics C-104.

Brunelle, Marc, Daryl Chow & Thụy Nhã Uyên Nguyễn. 2015. Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. In *Proceedings of ICPhS XVIII*. Glasgow.

Carroll, Lewis. 1866. *Alice's adventures in Wonderland*. New York: Appleton.
Ćavar, Małgorzata, Damir Cavar & Hilaria Cruz. 2016. Endangered language documentation: bootstrapping a Chatino speech corpus, forced aligner, ASR. In *Proceedings of LREC 2016, Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia.
Cruz, Emiliana. 2011. *Phonology, tone and the functions of tone in San Juan Quiahije Chatino*. Austin: University of Texas at Austin Ph.D. http://hdl.handle.net/2152/ETD-UT-2011-08-4280.
Cruz, Emiliana & Tony Woodbury. 2014. Finding a way into a family of tone languages: The story and methods of the Chatino Language Documentation Project. *Language Documentation and Conservation* 8. 490–524.
Dilley, L. & S. Shattuck-Hufnagel. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24. 423–444.
Do, Thi Ngoc Diep, Alexis Michaud & Eric Castelli. 2014. Towards the automatic processing of Yongning Na (Sino-Tibetan): developing a "light" acoustic model of the target language and testing "heavyweight" models from five national languages. In *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU 2014)*, 153–160. St Petersburg. http://halshs.archives-ouvertes.fr/halshs-00980431/.
Durand, Marguerite. 1930. *Etude sur les phonèmes postérieurs dans une articulation parisienne* (Petite Collection de l'intitut de Phonétique et Du Musée de La Parole et Du Geste). Paris: Didier.
Esch, Daan van, Ben Foley & Nay San. 2019. Future directions in technological support for language documentation. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1, 3.
Foley, Ben, Josh Arnold, Rolando Coto-Solano, Gautier Durantin & T. Mark Ellison. 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*, 200–204. Gurugram, India: ISCA.
Foley, Ben, Alina Rakhi, Nicholas Lambourne, Nicholas Buckeridge & Janet Wiles. 2019. Elpis, an accessible speech-to-text tool. In *Proceedings of Interspeech 2019*, 306–310. Graz.
Fougeron, Cécile. 1999. Prosodically conditioned articulatory variations: a review. *UCLA Working Papers in Phonetics* 97. 1–68.
Fougeron, Cécile. 2001. Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics* 29(2). 109–135.
Gendrot, Cédric. 2003. EGG and spectral slope investigation on final focalized positions in French. In *15th International Congress of Phonetic Sciences*. Barcelona.

Gomez-Marin, Alex. 2017. Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient? In *Decoding neural circuit structure and function*, 283–306. Springer.

Graves, Alex, Abdel-rahman Mohamed & Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649. IEEE.

Henderson, Eugénie J.A. 1985. Feature shuffling in Southeast Asian languages. In Suriya Ratanakul, David Thomas & Premsrirat Suwilai (eds.), *Southeast Asian Linguistic Studies presented to André-G. Haudricourt*, 1–22. Bangkok: Mahidol University.

Hjortnaes, Nils, Niko Partanen, Michael Rießler & Francis M. Tyers. 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*, 31–37. Wien: Association for Computational Linguistics.

Hohman, Fred, Andrew Head, Rich Caruana, Robert DeLine & Steven M. Drucker. 2019. Gamut: a design probe to understand how data scientists understand Machine Learning models. In. Glasgow.

Jacques, Guillaume & Alexis Michaud. 2011. Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze. *Diachronica* 28(4). 468–498.

Jiang, Zhengbao, Frank F. Xu, Jun Araki & Graham Neubig. 2019. How can we know what language models know? *arXiv:1911.12543 [cs]*. http://arxiv.org/abs/1911.12543 (9 February, 2020).

Jimerson, Robert & Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 4161–4166. Miyazaki.

Kahn, Jacob, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert & Christian Fuegen. 2019. Libri-Light: a benchmark for ASR with limited or no supervision. *arXiv preprint arXiv:1912.07875*.

Kuang, Jianjing. 2017. Creaky voice as a function of tonal categories and prosodic boundaries. In *Proceedings of Interspeech 2017*, 3216–3220. Stockholm.

Lapuschkin, Sebastian, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek & Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10(1). 1096.

Lidz, Liberty. 2010. *A descriptive grammar of Yongning Na (Mosuo)*. Austin: University of Texas, Department of linguistics Ph.D. https://repositories.lib.utexas.edu/bitstream/handle/2152/ETD-UT-2010-12-2643/LIDZ-DISSERTATION.pdf.

Lindblom, B. 1990. Explaining phonetic variation: a sketch of the H&H theory. In W.J. Hardcastle & Alain Marchal (eds.), *Speech production and speech modelling*, 403–439. Dordrecht: Kluwer.

Littell, Patrick, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox & Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, 2620–2632.

Martin, James H. & Daniel Jurafsky. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle, NJ: Pearson/Prentice Hall.

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

Michaud, Alexis. 2008. Phonemic and tonal analysis of Yongning Na. *Cahiers de linguistique - Asie Orientale* 37(2). 159–196.

Michaud, Alexis. 2012. Monosyllabicization: patterns of evolution in Asian languages. In Nicole Nau, Thomas Stolz & Cornelia Stroh (eds.), *Monosyllables: from phonology to typology*, 115–130. Berlin: Akademie Verlag. http://halshs.archives-ouvertes.fr/halshs-00436432/.

Michaud, Alexis. 2015. *Dictionnaire na-chinois-français*. https://halshs.archives-ouvertes.fr/halshs-01204645/.

Michaud, Alexis. 2017. *Tone in Yongning Na: lexical tones and morphotonology* (Studies in Diversity Linguistics 13). Berlin: Language Science Press. http://langsci-press.org/catalog/book/109.

Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig & 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.

Michaud, Alexis, Oliver Adams, Christopher Cox & Séverine Guillaume. 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, Alexis, Oliver Adams, Séverine Guillaume & Guillaume Wisniewski. 2020. Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo). In Christophe Benzitoun, Chloé Braud, Laurine Huber, David Langlois, Slim Ouni, Sylvain Pogodalla & Stéphane Schneider (eds.), *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, 451–462. Nancy, France: ATALA. https://hal.archives-ouvertes.fr/hal-02798572.

Michaud, Alexis, Séverine Guillaume, Guillaume Jacques, Đăng-Khoa Mạc, Michel Jacobson, Thu Hà Phạm & Matthew Deo. 2016. Contribuer au progrès solidaire des recherches et de la documentation : la Collection Pangloss et la Collection AuCo. In *Actes de la conférence conjointe JEP-TALN-RECITAL 2016, volume 1 : Journées d'Etude de la Parole*, vol. 1, 155–163. Paris: Association Francophone de la Communication Parlée. https://halshs.archives-ouvertes.fr/halshs-01341631/.

Michaud, Alexis, Andrew Hardie, Séverine Guillaume & Martine Toda. 2012. Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, Eric Castelli, Dong Minghui & Pham Thi Ngoc Yen, (eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)*, 169–172. Hanoi, Vietnam: MICA Institute, Hanoi University of Science and Technology.

Montavon, Grégoire, Wojciech Samek & Klaus-Robert Müller. 2017. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73. 1–15.

Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian & Petr Schwarz. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Pratap, Vineel, Awni Hannun, Qiantong Xu, Jeff Cai, Jacob Kahn, Gabriel Synnaeve, Vitaliy Liptchinsky & Ronan Collobert. 2018. wav2letter++: The fastest open-source speech recognition system. *arXiv:1812.07625 [cs]*. http://arxiv.org/abs/1812.07625 (9 February, 2020).

Rivière, Morgane, Armand Joulin, Pierre-Emmanuel Mazaré & Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. *arXiv preprint arXiv:2002.02848*.

Schneider, Steffen, Alexei Baevski, Ronan Collobert & Michael Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Shao, Bowei & Rachid Ridouane. 2018. La « voyelle apicale » en chinois de Jixi : caractéristiques acoustiques et comportement phonologique. In *XXXIIe Journées d'Études sur la Parole*, 685–693. ISCA. https://doi.org/10.21437/JEP.2018-78. http://www.isca-speech.org/archive/JEP_2018/abstracts/193446.html.

Shi, Tianze, Shun Kasahara, Teeraphon Pongkittiphan, Nobuaki Minematsu, Daisuke Saito & Keikichi Hirose. 2015. A measure of phonetic similarity to quantify pronunciation variation by using ASR technology. In *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: The University of Glasgow.

Sims, Michelle. 2010. Vowel space in the Athabaskan language of Tsuut'ina: Vowel shift in language attrition. In *Proceedings of the Conference on the Endangered Languages and Cultures of Native America*, vol. 1.

Sims, Michelle. 2011. Acoustic phonetic analysis as a means of defining the phonemic inventory: Evidence from the vowel space of Tsuut'ina. *Rochester Working Papers in the Language Sciences* 6. 1–18.

Soria, Claudia, Laurent Besacier & Laurette Pretorius. 2018. *Sustaining knowledge diversity in the digital age. Proceedings of the LREC 2018 Workshop "Collaboration and Computing for Under-Resourced Languages" (CCURL2018)*. Miyazaki: ELRA.

Stavness, Ian, Bryan Gick, Donald Derrick & Sidney Fels. 2012. Biomechanical modeling of English /r/ variants. *The Journal of the Acoustical Society of America* 131(5). EL355–EL360. https://doi.org/10.1121/1.3695407.

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

Tomashenko, Natalia & Yannick Estève. 2018. Impact des techniques d'adaptation au locuteur dans l'espace des paramètres pour des modèles acoustiques purement neuronaux. In *XXXIIe Journées d'Études sur la Parole*, 559–567. ISCA. https://doi.org/10.21437/JEP.2018-64. http://www.isca-speech.org/archive/JEP_2018/abstracts/192919.html (9 February, 2020).

Vaissière, Jacqueline. 1971. *Contribution à la synthèse par règles du français*. Grenoble.

Vaissière, Jacqueline. 2004. The perception of intonation. In David B. Pisoni & Robert E. Remez (eds.), *Handbook of Speech Perception* (Blackwell Textbooks in Linguistics), 236–263. Oxford, U.K. & Cambridge, Massachusetts: Blackwell.

Vaissière, Jacqueline. 2011a. On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. In *Proceedings of ICPhS XVII*. Hong Kong.

Vaissière, Jacqueline. 2011b. Proposals for a representation of sounds based on their main acoustico-perceptual properties. In Elizabeth Hume, John Goldsmith & W. Leo Wetzels (eds.), *Tones and Features*, 306–330. Berlin: De Gruyter Mouton.

Vasile, Aurelia, Séverine Guillaume, Mourad Aouini & Alexis Michaud. 2020. Le Digital Object Identifier, une impérieuse nécessité ? L'exemple de l'attribution de DOI à la Collection Pangloss, archive ouverte de langues en danger. *I2D - Information, données & documents*. https://halshs.archives-ouvertes.fr/halshs-02870206.

Watanabe, Shinji, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner & Nanxin Chen. 2018. ESPnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*.

Wisniewski, Guillaume, Séverine Guillaume & Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In Dorothee Beermann, Laurent Besacier, Sakriani Sakti & Claudia 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*, 306–315. Marseille, France: European Language Resources Association (ELRA).

Wu, Minghao, Fei Liu & Trevor Cohn. 2018. Evaluating the utility of hand-crafted features in sequence labelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2850–2856. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1310. http://aclweb.org/anthology/D18-1310 (9 February, 2020).

ABSTRACT. - Automatic speech recognition systems now achieve high levels of accuracy with relatively small amounts of training data: on the order two to three hours of transcribed speech, instead of tens of hours for previous tools. Beyond the practical usefulness of these technological advances for linguistic documentation tasks, use of automatic transcription also yields some linguistic insights. Acoustic models are built on the basis of the linguist's transcriptions, and thus encapsulate linguistic hypotheses and assumptions. To what extent can acoustic models be examined in turn by the linguist? What can we learn from this renewed confrontation with the acoustic signal? The present study is based on examples from the Native language (Sino-Tibetan family) to illustrate how error analysis allows a renewed confrontation with the data. Among other benefits, error analysis allows for a renewed exploration of phonetic detail: examining the output of phonemic transcription software compared with spectrographic and aural evidence. Some reflections on experiments of automatic transcription of the Tsuut'ina language (Dene family) are also presented.

摘要. - 目前，自动语音识别系统使用相对较少的训练数据就能达到很高的准确度：以前需要几十个小时才能完成的语音转录任务现在只需两三个小时即可完成。除了技术进步对语言记录任务的实际效率作用外，使用自动转录也产生了一些新的语言学观点：声学模型是建立在语言学家的转录基础上的，因此也涵盖了语言学的假设和假定。声学模型在多大程度上可以被语言学家用来进行反证和考察？我们能从这种对声学信号的重新面对中学习到什么？本研究基于纳语（摩梭话）的例子来说明误差分析是如何让我们重新面对数据的。除其他优势以外，误差分析还可以重新探索语音细节：将音位转录软件的输出与频谱和听觉证据进行对比研究。还提出了对北美大陆德内语支（阿萨巴斯卡语支）语言自动转录实验的一些思考。