Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Analyse d'erreurs de transcriptions phonémiques automatiques d'une langue « rare » : le na (mosuo)

Alexis Michaud[1] Oliver Adams[2] Séverine Guillaume[1] Guillaume Wisniewski[3]

(1) Langues et Civilisations à Tradition Orale (LACITO), CNRS – Université Sorbonne Nouvelle – INALCO, 7 rue Guy Môquet, 94800 Villejuif, France
(2) Miner & Kasch, 8174 Lark Brown Rd #101, Elkridge, MD 21075, Etats-Unis d'Amérique
(3) Laboratoire de Linguistique Formelle (LLF), CNRS – Université Paris-Diderot, Case 7031, 5 rue Thomas Mann, 75013 Paris, France

alexis.michaud@cnrs.fr, oliver.adams@gmail.com, severine.guillaume@cnrs.fr, guillaume.wisniewski@linguist.univ-paris-diderot.fr

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-delà de l'intérêt pratique que présentent ces avancées technologiques pour les tâches de documentation de langues rares et en danger, se pose la question de leur apport pour la réflexion du phonéticien/phonologue. En effet, le modèle acoustique prend en entrée des transcriptions qui reposent sur un ensemble d'hypothèses 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 ? Notre étude s'appuie sur des exemples d'une langue « rare » de la famille sino-tibétaine, le na (mosuo), pour illustrer la façon dont l'analyse d'erreurs permet une confrontation renouvelée avec le signal acoustique.

ABSTRACT

Analyzing errors in automatic phonemic transcriptions of the Na (Mosuo) language (Sino-Tibetan family)

Automatic phonemic transcription tools now reach high levels of accuracy on a single speaker with relatively small amounts of training data: on the order two to three hours of transcribed speech. Beyond its practical usefulness for language documentation, use of automatic transcription also yields some insights for phoneticians/phonologists. Acoustic models are built on the basis of the linguist's transcriptions, and thus encapsulate linguistic hypotheses and assumptions. To what extent can the acoustic model be examined by the linguist? The present report explores this topic by going into qualitative error analysis on Yongning Na (Sino-Tibetan). 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.

MOTS-CLÉS : transcription phonologique, reconnaissance de la parole, apprentissage machine, analyse d'erreurs, interdisciplinarité, documentation linguistique assistée par ordinateur.

KEYWORDS: phonological transcription, speech recognition, machine learning, error analysis, interdisciplinarity, Computational Language Documentation.

Volume 1 : Journées d'Études sur la Parole, pages 451–462.
HAL : hal-02798572.

1 Introduction[1]

1.1 Phonétique et reconnaissance automatique de la parole

La reconnaissance vocale 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, dits d'*intelligence artificielle*. Dans ce contexte, il ne paraît pas inutile de souligner que le dialogue interdisciplinaire demeure aussi pertinent que jamais à l'ère de l'apprentissage machine. Informaticiens et linguistes ont tout intérêt à collaborer afin de concevoir et déployer des outils innovants, et d'en tirer parti pour la recherche (Adda et al., 2016; Neubig et al., 2020). En effet, les collaborations entre linguistes et spécialistes du Traitement Automatique des Langues Naturelles ne sont pas seulement utiles à une meilleure efficacité pratique des outils (l'apprentissage statistique supervisé, qui repose sur une annotation raisonnée, donne de meilleurs résultats que l'apprentissage non supervisé : Jimerson & Prud'hommeaux, 2018; Wu et al., 2018) : elles sont en outre prometteuses pour les sciences de la parole. Historiquement, des collaborations entre linguistes et chercheurs en intelligence artificielle se sont attachées à mieux cerner la nature du phonème (question-clef de la phonétique/phonologie : voir notamment Jones, 1950) et à examiner la possibilité de le représenter par des modèles acoustiques statistiques. Entre autres et nombreux travaux, on citera les études qui s'appuient sur deux grands corpus d'anglais américain, « Texas Instruments/Massachusetts Institute of Technology », TIMIT (Garofolo et al., 1993) et SWITCHBOARD (Godfrey et al., 1992), pour mener une analyse des erreurs de reconnaissance automatique par des systèmes utilisant des modèles acoustiques statistiques. Le premier constat est celui d'une distance considérable entre les réalisations canoniques des phonèmes et la parole spontanée : « De nombreux mots sont articulés d'une manière qui omet ou transforme profondément les propriétés phonétiques des phonèmes qui les constituent, ce qui entraîne une grande variabilité dans la prononciation d'un même mot. Souvent, on ne trouve d'un segment qu'un indice ténu, et ce bien que le signal soit tout à fait intelligible »[2] (Greenberg et al., 1996, p. 24). La présente communication se veut une (modeste) poursuite de ces réflexions. Elle repose sur l'utilisation d'un outil de transcription phonémique automatique, le logiciel `Persephone`. Le développement de cet outil est appelé à déboucher, à moyen terme (d'ici quelques années), sur un logiciel complet de reconnaissance de la parole, qui reconnaisse des mots entiers, puis des phrases entières (à titre d'exemple, voir les travaux de Hjortnaes et al., 2020 concernant le komi, langue ouralienne). Le stade actuel, celui d'une transcription phonémique, présente l'intérêt de demeurer au plus proche du signal acoustique, de sorte que l'analyse d'erreurs permet une confrontation détaillée avec le signal acoustique.

---

1 Le présent exposé reprend des résultats présentés dans une communication (en anglais) au XIX[e] Congrès des sciences phonétiques (Michaud et al., 2019). Cette communication portait non seulement sur la langue na, mais aussi sur le tsuut'ina, langue de la famille dene (athabasque), parlée dans l'ouest du Canada. Faute de place, les résultats sur les données tsuut'ina ne peuvent être exposés ici. Le lecteur intéressé est renvoyé au texte anglais, ainsi qu'à la présentation en vidéo du *plugin* créé par Christopher Cox pour faciliter l'emploi de `Persephone` par ses collaborateurs tsuut'ina (https://www.youtube.com/watch?v=-pDOEqRpZKs).

2 *Texte original :* Many words are articulated in such a fashion as to either omit or significantly transform the phonetic properties of phonemic constituents, thus resulting in wide variation of word pronunciations. Often, only the barest hint of a segment is realized phonetically, in spite of good intelligibility.

1.2 Genèse de l'outil `Persephone` et perspectives de développement

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 et al., 2019). L'outil logiciel `Persephone`, disponible en ligne (https://github.com/persephone-tools/persephone) sous licence libre, est issu de recherches exploratoires menées par Oliver Adams au fil de son travail de thèse (Adams, 2017). La façon dont notre collaboration s'est nouée est détaillée dans un article paru dans une revue spécialisée dans les questions de documentation et conservation de langues en danger (Michaud et al., 2018). Deux communications à des colloques de Traitement Automatique des Langues exposent le fonctionnement de l'outil (Adams et al., 2017, 2018).

Au plan du développement logiciel, notre projet actuel (2020-2023) consiste à contribuer à l'élaboration d'une interface utilisateur unique (Foley et al., 2019) 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`, mais aussi `wav2letter++` (Pratap et al., 2018), `KALDI`, `ESPnet` (Watanabe et al., 2018)... 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, par exemple. Dans le présent travail, l'accent n'est pas mis sur ces questions, mais sur les possibilités qu'offre la transcription automatique pour la recherche phonétique.

2 Méthode

2.1 Corpus employé

Les données employées, intégralement disponible en ligne, sont celles d'un corpus réuni au fil d'enquêtes linguistiques sur le terrain (Michaud et al., 2012) au sujet du na, langue sino-tibétaine parlée à la frontière entre les provinces chinoises du Yunnan et du Sichuan (Lidz, 2010). La phonotactique de la langue na est relativement simple : chaque syllabe comporte une consonne initiale (l'inventaire consonantique est présenté dans le tableau 1), l'une des voyelles (rimes) suivantes : /i e æ ɑ ɯ u ɤ o ɣ ɻ wæ wɑ wɤ jæ jɤ jo/ (pour plus de précisions : Michaud, 2008, 2017, pp. 447–486), et un ton. 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). Il existe cinq tons en « phonologie de surface » : Haut, Moyen, Bas, Bas-montant, et Moyen-montant (noté ˥, ˧, ˩, ˧˥, ˥˧).

	bilabiales	dentales	alvéolo-palatales	rétroflexes	vélaires	uvulaires	glottales
occlusives	pʰ p b	tʰ t d		ʈʰ ʈ ɖ	kʰ k g	qʰ q	ʔ
affriquées		tsʰ ts dz	tɕʰ tɕ dʑ	ʈʂʰ ʈʂ ɖʐ			
nasales	m	n	ɲ	ɳ	ŋ		
fricatives		s z	ɕ ʑ	ʂ ʐ		ʁ	h
latérales		ɬ l					
approximante				ɻ			

TABLEAU 1 : Les consonnes du na de Yongning.

2.2 Les algorithmes : principes de fonctionnement de l'outil `Persephone`

Le logiciel `Persephone` appartient à la génération 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, p. 561). Cette propriété du modèle permet de traiter, outre les phonèmes, l'information non segmentale, telle que les tons lexicaux (et tout autre type d'événement figuré dans la transcription fournie en entrée, par exemple un découpage en mots prosodiques).

Dans les expériences relatées ici, l'entraînement du modèle acoustique s'effectue à partir de zéro, sans recourir à des modèles déjà initialisés à partir de données multilingues. Le modèle acoustique est entraîné sur les données d'une unique locutrice. Cela limite d'emblée la portée des généralisations, du fait qu'il n'est pas possible de faire la part des habitudes spécifiques à la locutrice en question. Mais à l'inverse, le choix d'un corpus mono-locuteur permet d'exclure un important facteur de variabilité, et ainsi de pouvoir tirer des conclusions plus certaines, en attendant le stade (évidemment prévu pour la suite) d'une généralisation des observations réalisées. Au plan technique, une expérimentation systématique sur sept langues de la Collection Pangloss (Wisniewski et al., 2020) établit clairement que le passage d'un mode mono-locuteur à un mode multi-locuteurs demandera une amélioration des outils logiciels (selon les méthodes exposées par Tomashenko et al., 2020; Tomashenko & Estève, 2018).

Pour plus d'informations, on renverra à la documentation disponible en ligne[3]. On signalera également des exposés en vidéo (en anglais) au sujet de `Persephone`[4] et de son intégration dans le logiciel de documentation linguistique `ELAN`[5].

[3] https://persephone.readthedocs.io/en/stable/
[4] https://www.youtube.com/watch?v=IwWKqxQ7Qng
[5] https://www.youtube.com/watch?v=-pDOEqRpZKs

2.3 La validation croisée

La méthode employée est la *validation croisée*. L'un des ving-sept documents est retranché du corpus, 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 na. La validation croisée est une méthode pour éviter des biais d'apprentissage des modèles : il s'agit d'éviter que le modèle ne soit testé sur des données qui faisaient partie du corpus d'apprentissage (erreur classique *de débutant* dans le domaine de l'apprentissage machine). La validation croisée est particulièrement utile dans le cas où le corpus est petit, comme dans le scénario qui nous intéresse ici : cette méthode permet d'utiliser au mieux les données, en les employant comme données de test dans une des expériences, tout en les conservant parmi les données d'entraînement pour les autres.

2.4 Comparaison entre transcriptions générées automatiquement et transcriptions manuelles : le 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 à des taux d'erreur de l'ordre de 17% pour la langue na (Adams et al., 2018), en très net progrès par rapport à une étude-pilote réalisée sur les mêmes données au moyen de `CMU-Sphinx` (Do et al., 2014). Au-delà de ce résultat général encourageant, nous avons choisi d'examiner des exemples détaillés, l'un après l'autre, plutôt que d'aborder l'analyse d'erreurs au moyen d'outils statistiques. 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 vingt-sept documents PDF sont disponibles en ligne[6]. Les documents du corpus d'entraînement peuvent être consultés dans la Collection Pangloss (Michaud et al., 2016)[7]. Nous sommes conscients du fait que les observations présentées ici ne constituent qu'une première approche, qu'il sera utile de poursuivre par une analyse statistique dans les règles de l'art.

3 Premières observations

3.1 La situation particulière des noms propres quadrisyllabiques

Un exemple de transcription automatique suivi de la transcription de référence (manuelle) en vis-à-vis est fourni ci-dessous. Les gloses figurent en exemple (1).

æ̃˩	ʈʂʰe˦	ɖɯ˦	mæ˩		ʈʂʰɯ˦	bi˦	mæ˩	pi˦	dʑo˩	*transcription automatique*
ɻ̩˩	ʈʂʰe˦	ɖɯ˩	mɑ˩	ɻ̩˩	ʈʂʰe˦	ɖɯ˩	mɑ˩	pi˦	dʑo˩	*transcription manuelle*

[6] https://github.com/alexis-michaud/na/tree/master/Persephone/2018_08_StoryFoldCrossValidation
[7] https://pangloss.cnrs.fr/corpus/list_rsc.php?lg=Na&name=na

(1) ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩ pi˧ dʑo˩
Erchei-Ddeema *(nom propre)* dire TOPICALISATEUR
Elle a crié : « Erchei-Ddeema ! Erchei-Ddeema ! » (Texte : *Enterrée vive*, phrase 13. DOI : 10.24397/pangloss-0004537#S13)

Dans cet énoncé, on relève des erreurs de transcription sur les deux occurrences du nom Erchei-Ddeema (nom d'un des principaux protagonistes). La forme phonémique de ce nom est /ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩/. Au vu du taux d'erreur globalement faible (de l'ordre de 17%), il est frappant d'observer neuf erreurs en l'espace d'à peine huit syllabes. L'examen des onze occurrences de ce nom propre dans le texte (reproduites ci-dessous) révèle qu'aucune n'est exempte d'erreurs.

p æ ˩ ʈʂʰ ɯ ˥ ɖ ɯ ˧ m ɣ˩	æ̃˩ ʈʂʰ e ˧ ɖ ɯ ˧ m æ ˩	∅ ʈʂʰ ɯ ˧ b i ˧ m æ ˩
ɑ ˩ ʈʂʰ e ˧ ɖ ɯ ˧ m ɤ ˩	∅ tʰ i ˥ ɖ ɯ ˧ m ɑ ˩	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧
ɻ˩ ʈʂʰ e ˧ ɖ ɯ ˧ m ɤ ˧	ɻ˧ ʈʂʰ ɯ ˧ dʑ ɯ ˧ m ɤ ˧	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧
ɻ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧	æ ˩ ʈʂʰ ɯ ˧ ɖ ɯ ˧ m ɤ ˧	

TABLE 2 : Transcription automatique des onze occurrences du nom propre Erchei-Ddeema /ɻ˩ ʈʂʰe˧ ɖɯ˩ ma˩/. Le symbole de l'ensemble vide ∅ indique une syllabe manquante.

La première syllabe, l'approximante syllabique / ɻ/, est identifiée comme une voyelle dans six cas, et manque tout à fait dans deux cas. 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 autres rimes. 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 exemple en est fourni en Figure 1. 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.

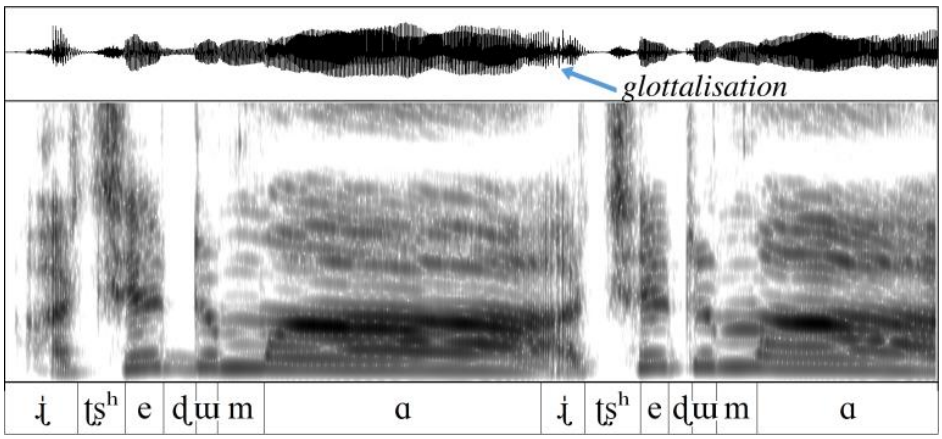

FIGURE 3 : Spectrogramme correspondant à l'exemple (1).

La voyelle de la seconde syllabe du nom /ɻ˩ 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 [ʈʂʰz̩] (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 [ʈʂʰ].

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 spectaculaire 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 largement 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 donc 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[8].

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). Par exemple, 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 (Brunelle et al., 2015) : une préposition vietnamienne homophone d'un verbe en diffère moins au plan phonétique que ne le laisserait attendre l'exemple des langues comme le français ou l'anglais. 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. 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

[8] Ce raisonnement ne vaut pas pour les mots courts fréquents et fortement prédictibles dans le discours (tels que les mots outils/grammaticaux), souvent hypo-articulés.

grammatical du mot (classe morphosyntaxique) et sa fréquence. On peut donc espérer que la poursuite de ces observations apporte une contribution de nature typologique aux questions de variation allophonique et de « prosodie articulatoire » (Fougeron, 1999, 2001).

3.2 Interprétation des observations

Les quadri-syllabes ne sont pas très fréquents en na (et de manière générale, peu fréquents dans les corpus oraux, dans de nombreuses langues). Les observations rapportées ci-dessus au sujet du nom propre /ɻ˩ ʈʂʰe˧ ɖɯ˩ mɑ˩/ éclairent une des limites du système automatique : le biais statistique qui conduit à accorder plus de poids aux phénomènes plus fréquent, avec pour résultat de moins bonnes performances pour une catégorie qui est marginale en terme de fréquence dans le corpus d'apprentissage.

4 Conclusion et perspectives

Les travaux présentés ici n'en sont qu'à leurs débuts, mais 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. Il existe diverses méthodes pour explorer ce domaine (Hohman et al., 2019; Jiang et al., 2019; Lapuschkin et al., 2019; Montavon et al., 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. 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, 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.

Remerciements

Nos vifs remerciements aux locuteurs et amis na. Nous remercions vivement les deux relecteurs des *Journées d'Étude de la Parole*, ainsi que tous les collègues qui participent au développement et à l'utilisation d'outils de reconnaissance automatique pour langues peu dotées ; qu'ils nous pardonnent de ne pas nous livrer à l'exercice impossible qui consisterait à dresser une liste un tant soit peu complète.

Le logiciel `Persephone` a bénéficié en 2018-2019 du soutien de l'Université du Queensland et d'une bourse d'innovation transdisciplinaire du *Centre of Excellence for the Dynamics of Language* du Conseil australien de la recherche (ARC). Il bénéficie actuellement du soutien du projet franco-allemand « La documentation computationnelle des langues à l'horizon 2025 » (CLD 2025, ANR-19-CE38-0015-04) et du projet d'Institut des Langues Rares (ILARA) de l'École pratique des Hautes Études (dans le cadre du plan *Sciences humaines et sociales 2020* du Ministère de

l'enseignement supérieur, de la recherche et de l'innovation). Le présent travail s'inscrit en outre dans le cadre du Labex « Fondements empiriques de la linguistique » (EFL, ANR-10-LABX-0083).

Références

ADAMS, O. (2017). *Automatic understanding of unwritten languages* [Ph.D.]. The University of Melbourne.

ADAMS, O., COHN, T., NEUBIG, G., CRUZ, H., BIRD, S., & MICHAUD, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 3356–3365. HAL : https://halshs.archives-ouvertes.fr/halshs-01709648

ADAMS, O., COHN, T., NEUBIG, G., & MICHAUD, A. (2017). Phonemic transcription of low-resource tonal languages. *Proceedings of ALTA 2017 (Australasian Language Technology Association Workshop)*, 53–60. HAL : https://halshs.archives-ouvertes.fr/halshs-01656683

ADDA, G., STÜKER, S., ADDA-DECKER, M., AMBOUROUE, O., BESACIER, L., BLACHON, D., BONNEAU-MAYNARD, H., GODARD, P., HAMLAOUI, F., IDIATOV, D., KOUARATA, G.-N., LAMEL, L., MAKASSO, E.-M., RIALLAND, A., VAN DE VELDE, M., YVON, F., & ZERBIAN, S. (2016). Breaking the unwritten language barrier: The BULB Project. *SLTU-2016 5th Workshop on Spoken Language Technologies for Under-Resourced Languages 09-12 May 2016 Yogyakarta, Indonesia*, *81*(Supplement C), 8–14. DOI : 10.1016/j.procs.2016.04.023

BRUNELLE, M., CHOW, D., & NGUYỄN, T. N. U. (2015). Effects of lexical frequency and lexical category on the duration of Vietnamese syllables. *Proceedings of ICPhS XVIII*. International Congress of the Phonetic Sciences XVIII, Glasgow.

DILLEY, L., & SHATTUCK-HUFNAGEL, S. (1996). Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics*, *24*, 423–444.

DO, T. N. D., MICHAUD, A., & CASTELLI, E. (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. *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, 153–160. HAL : http://halshs.archives-ouvertes.fr/halshs-00980431

DURAND, M. (1930). *Etude sur les phonèmes postérieurs dans une articulation parisienne*. Didier.

FOLEY, B., ARNOLD, J., COTO-SOLANO, R., DURANTIN, G., & ELLISON, T. M. (2018). Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). *Proceedings of the 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), 29-31 August 2018*, 200–204.

FOLEY, B., RAKHI, A., LAMBOURNE, N., BUCKERIDGE, N., & WILES, J. (2019). Elpis, an accessible speech-to-text tool. *Proceedings of Interspeech 2019*, 306–310.

FOUGERON, C. (1999). Prosodically conditioned articulatory variations: A review. *UCLA Working Papers in Phonetics*, *97*, 1–68.

FOUGERON, C. (2001). Articulatory properties of initial segments in several prosodic constituents in French. *Journal of Phonetics*, *29*(2), 109–135.

GAROFOLO, J. S., LAMEL, L. F., FISHER, W. M., FISCUS, J. G., & PALLETT, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report n*, *93*.

GODFREY, J. J., HOLLIMAN, E. C., & MCDANIEL, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 517–520.

GOMEZ-MARIN, A. (2017). Causal circuit explanations of behavior: Are necessity and sufficiency necessary and sufficient? In *Decoding neural circuit structure and function* (pp. 283–306). Springer.

GRAVES, A., MOHAMED, A., & HINTON, G. (2013). Speech recognition with deep recurrent neural networks. *ICASSP 2013 - 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 6645–6649.

GREENBERG, S., HOLLENBACK, J., & ELLIS, D. (1996). Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of the International Conference on Spoken Language Processing*, *96*, 24–27.

HJORTNAES, N., PARTANEN, N., RIEßLER, M., & TYERS, F. M. (2020). Towards a speech recognizer for Komi, an endangered and low-resource Uralic language. *Proceedings of the Sixth International Workshop on Computational Linguistics of Uralic Languages*, 31–37.

HOHMAN, F., HEAD, A., CARUANA, R., DELINE, R., & DRUCKER, S. M. (2019). *Gamut: A design probe to understand how data scientists understand Machine Learning models*. ACM CHI Conference on Human Factors in Computing Systems, Glasgow.

JACQUES, G., & MICHAUD, A. (2011). Approaching the historical phonology of three highly eroded Sino-Tibetan languages: Naxi, Na and Laze. *Diachronica*, *28*(4), 468–498.

JIANG, Z., XU, F. F., ARAKI, J., & NEUBIG, G. (2019). How can we know what language models know? *ArXiv:1911.12543*. http://arxiv.org/abs/1911.12543

JIMERSON, R., & PRUD'HOMMEAUX, E. (2018). ASR for documenting acutely under-resourced indigenous languages. *Proceedings of LREC 2018 (Language Resources and Evaluation Conference)*, 4161–4166.

JONES, D. (1950). *The Phoneme, its Nature and Use*. Heffer.

KUANG, J. (2017). Creaky voice as a function of tonal categories and prosodic boundaries. *Proceedings of Interspeech 2017*, 3216–3220.

LAPUSCHKIN, S., WÄLDCHEN, S., BINDER, A., MONTAVON, G., SAMEK, W., & MÜLLER, K.-R. (2019). Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, *10*(1), 1096.

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

LITTELL, P., KAZANTSEVA, A., KUHN, R., PINE, A., ARPPE, A., COX, C., & JUNKER, M.-O. (2018). Indigenous language technologies in Canada: Assessment, challenges, and successes. *Proceedings of the 27th International Conference on Computational Linguistics*, 2620–2632.

MICHAILOVSKY, B., MAZAUDON, M., MICHAUD, A., GUILLAUME, S., FRANÇOIS, A., & ADAMOU, E. (2014). Documenting and researching endangered languages: The Pangloss Collection. *Language Documentation and Conservation*, *8*, 119–135. HAL : https://halshs.archives-ouvertes.fr/halshs-01003734

MICHAUD, A. (2008). Phonemic and tonal analysis of Yongning Na. *Cahiers de Linguistique - Asie Orientale*, *37*(2), 159–196. HAL : https://halshs.archives-ouvertes.fr/halshs-00358610

MICHAUD, A. (2012). Monosyllabicization: Patterns of evolution in Asian languages. In N. Nau, T. Stolz, & C. Stroh (Eds.), *Monosyllables: From phonology to typology* (pp. 115–130). Akademie Verlag. HAL : http://halshs.archives-ouvertes.fr/halshs-00436432

MICHAUD, A. (2015). *Dictionnaire na-chinois-français*. HAL : https://halshs.archives-ouvertes.fr/halshs-01204645

MICHAUD, A. (2017). *Tone in Yongning Na: Lexical tones and morphotonology*. Language Science Press. http://langsci-press.org/catalog/book/109

MICHAUD, A., ADAMS, O., COHN, T., NEUBIG, G., & GUILLAUME, S. (2018). Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation and Conservation*, *12*, 393–429. HAL : https://halshs.archives-ouvertes.fr/halshs-01841979

MICHAUD, A., ADAMS, O., COX, C., & GUILLAUME, S. (2019). Phonetic lessons from automatic phonemic transcription: Preliminary reflections on Na (Sino-Tibetan) and Tsuut'ina (Dene) data. *Proceedings of ICPhS XIX (19th International Congress of Phonetic Sciences)*. ICPhS XIX (19th International Congress of Phonetic Sciences), Melbourne. HAL : https://halshs.archives-ouvertes.fr/halshs-02059313

MICHAUD, A., GUILLAUME, S., JACQUES, G., MẠC, Đ.-K., JACOBSON, M., PHẠM, T. H., & DEO, M. (2016). Contribuer au progrès solidaire des recherches et de la documentation: La Collection Pangloss et la Collection AuCo. *Actes de La Conférence Conjointe JEP-TALN-RECITAL 2016, Volume 1 : Journées d'Etude de La Parole*, *1*, 155–163. HAL : https://halshs.archives-ouvertes.fr/halshs-01341631

MICHAUD, A., HARDIE, A., GUILLAUME, S., & TODA, M. (2012). Combining documentation and research: Ongoing work on an endangered language. In Xiong Deyi, E. Castelli, Dong Minghui, & Pham Thi Ngoc Yen, (Eds.), *Proceedings of IALP 2012 (2012 International Conference on Asian Language Processing)* (pp. 169–172). MICA Institute, Hanoi University of Science and Technology. HAL : https://halshs.archives-ouvertes.fr/halshs-00731261

MONTAVON, G., SAMEK, W., & MÜLLER, K.-R. (2017). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15.

NEUBIG, G., RIJHWANI, S., PALMER, A., MACKENZIE, J., CRUZ, H., LI, X., LEE, M., CHAUDHARY, A., GESSLER, L., & ABNEY, S. (2020). A summary of the first Workshop on Language Technology for Language Documentation and Revitalization. *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-Resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*. Marseille, France. *ArXiv:2004.13203 [Cs]*. https://arxiv.org/abs/2004.13203

PRATAP, V., HANNUN, A., XU, Q., CAI, J., KAHN, J., SYNNAEVE, G., LIPTCHINSKY, V., & COLLOBERT, R. (2018). wav2letter++: The fastest open-source speech recognition system. *ArXiv:1812.07625 [Cs]*. http://arxiv.org/abs/1812.07625

SHAO, B., & RIDOUANE, R. (2018). La « voyelle apicale » en chinois de Jixi: Caractéristiques acoustiques et comportement phonologique. *XXXIIe Journées d'Études Sur La Parole*, 685–693. DOI : 10.21437/JEP.2018-78

STAVNESS, I., GICK, B., DERRICK, D., & FELS, S. (2012). Biomechanical modeling of English /r/ variants. *The Journal of the Acoustical Society of America*, *131*(5), EL355–EL360. DOI : 10.1121/1.3695407

THIEBERGER, N. (2017). LD&C possibilities for the next decade. *Language Documentation and Conservation*, *11*, 1–4.

TOMASHENKO, N., & ESTÈVE, Y. (2018). Impact des techniques d'adaptation au locuteur dans l'espace des paramètres pour des modèles acoustiques purement neuronaux. *XXXIIe Journées d'Études Sur La Parole*, 559–567. DOI : 10.21437/JEP.2018-64

TOMASHENKO, N., KHOKHLOV, Y., & ESTÈVE, Y. (2020). *Exploring Gaussian mixture model framework for speaker adaptation of deep neural network acoustic models*. HAL : https://hal.archives-ouvertes.fr/hal-02551714

VAISSIÈRE, J. (2004). The perception of intonation. In D. B. PISONI & R. E. REMEZ (Eds.), *Handbook of Speech Perception* (pp. 236–263). Blackwell.

VAISSIÈRE, J. (2011a). On the acoustic and perceptual characterization of reference vowels in a cross-language perspective. *Proceedings of ICPhS XVII*. ICPhS XVII, Hong Kong.

VAISSIÈRE, J. (2011b). Proposals for a representation of sounds based on their main acoustico-perceptual properties. In E. HUME, J. GOLDSMITH, & W. L. WETZELS (Eds.), *Tones and Features* (pp. 306–330). De Gruyter Mouton.

VAN ESCH, D., FOLEY, B., & SAN, N. (2019). Future directions in technological support for language documentation. *Proceedings of the Workshop on Computational Methods for Endangered Languages*, *1*, 3. https://www.aclweb.org/anthology/W19-6003.pdf

WATANABE, S., HORI, T., KARITA, S., HAYASHI, T., NISHITOBA, J., UNNO, Y., SOPLIN, N. E. Y., HEYMANN, J., WIESNER, M., & CHEN, N. (2018). Espnet: End-to-end speech processing toolkit. ArXiv:1804.00015.

WISNIEWSKI, G., GUILLAUME, S., & MICHAUD, A. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? *Proceedings of the 1st Joint SLTU (Spoken Language Technologies for Under-Resourced Languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop*. 1st Joint SLTU (Spoken Language Technologies for Under-resourced languages) and CCURL (Collaboration and Computing for Under-Resourced Languages) Workshop, Marseille, France. HAL : https://halshs.archives-ouvertes.fr/hal-02513914

WU, M., LIU, F., & COHN, T. (2018). Evaluating the utility of hand-crafted features in sequence labelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2850–2856. DOI : 10.18653/v1/D18-1310