HAL
open science

Construction faiblement supervisée d'un phonétiseur pour la langue Iban à partir de ressources en Malais

Sarah Samson Juan, Laurent Besacier, Solange Rossato

To cite this version:

Sarah Samson Juan, Laurent Besacier, Solange Rossato. Construction faiblement supervisée d'un phonétiseur pour la langue Iban à partir de ressources en Malais. Journées d'Etude sur la Parole (JEP), Jun 2014, Le Mans, France. hal-01002921

HAL Id: hal-01002921

https://inria.hal.science/hal-01002921

Submitted on 7 Jun 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Construction faiblement supervisée d'un phonétiseur pour la langue Iban à partir de ressources en Malais

Sarah Samson Juan[1], Laurent Besacier[1], Solange Rossato[1]
(1) Grenoble Informatics Laboratory (LIG)
University Grenoble-Alpes
Grenoble, France
(sarah.samson-juan@imag.fr, laurent.besacier@imag.fr, solange.rossato@imag.fr

RÉSUMÉ

Cet article décrit notre collecte de ressources pour la langue iban (parlée notamment sur l'île de Bornéo), dans l'objectif de construire un système de reconnaissance automatique de la parole pour cette langue. Nous nous sommes plus particulièrement focalisés sur une méthodologie d'amorçage du lexique phonétisé à partir d'une langue proche (le malais). Les performances des premiers systèmes de reconnaissance automatique de la parole construits pour l'iban ($< 20\%$ WER) montrent que l'utilisation d'un phonétiseur déjà disponible dans une langue proche (le malais) est une option tout à fait viable pour amorcer le développement d'un système de RAP dans une nouvelle langue très peu dotée. Une première analyse des erreurs fait ressortir des problèmes bien connus pour les langues peu dotées : problèmes de normalisation de l'orthographe, erreurs liées à la morphologie (séparation ou non des affixes de la racine).

ABSTRACT

Weakly supervised G2P training for Iban language using Malay Ressources

This paper describes our experiments and results on using a local dominant language in Malaysia (Malay), to bootstrap automatic speech recognition (ASR) for a very under-resourced language : iban (also spoken in Malaysia on the Borneo Island part). Resources in iban for building a speech recognition were nonexistent. For this, we tried to take advantage of a language from the same family with several similarities. First, to deal with the pronunciation dictionary, we proposed a bootstrapping strategy to develop an iban pronunciation lexicon from a Malay one. A hybrid version, mix of Malay and iban pronunciations, was also built and evaluated. Following this, we experimented with three iban ASRs ; each depended on either one of the three different pronunciation dictionaries : Malay, iban or hybrid.

MOTS-CLÉS : langues peu-dotées, reconnaissance automatique de la parole, conversion graphème-phonème, iban, malais, amorçage, Kaldi.

KEYWORDS: under-resourced languages, speech recognition, G2P conversion, iban language, Malay language, bootstrapping, Kaldi.

1 Introduction

Le dictionnaire de prononciation est une ressource cruciale pour certaines technologies vocales telles que la reconnaissance automatique de la parole (RAP) ou la synthèse de la parole (TTS). Construire un lexique phonétisé pour une nouvelle langue et pour plusieurs dizaines de milliers de mots peut prendre un temps considérable. On peut bien sûr envisager de créer un système

automatique de conversion graphème-phonème (que nous nommerons, dans la suite de cette article G2P pour *Grapheme-to-Phoneme Conversion*). Cependant, la construction d'un système G2P nécessite des règles de phonétisation formalisées (méthode à base de règles) ou un corpus d'apprentissage contenant des mots déjà phonétisés (méthode empirique). Cet article propose d'utiliser des ressources déjà disponibles dans une langue proche assez bien dotée (le malais), pour amorcer la construction d'un système de conversion graphème-phonème pour une langue très peu dotée (l'iban).

Un amorçage de système G2P a déjà été proposé pour aider à la création de lexiques de prononciation pour les langues telles que l'afrikaans et le népalais ((Davel et Martirosian, 2009), (Maskey *et al.*, 2004)). Dans ces approches, un petit lexique disponible dans la langue visée était utilisé pour construire un premier système de conversion graphème-phonème. Ce système était ensuite utilisé pour générer les prononciations d'un plus gros vocabulaire et ses sorties étaient vérifiées et (éventuellement) corrigées par un locuteur natif de la langue cible. Le nouveau dictionnaire de prononciation, de plus grande taille, était ensuite utilisé pour construire le système G2P final. Cependant, pour de nombreuses langues pauvrement dotées (Besacier *et al.*, 2014), même un petit lexique initial est difficile (voir impossible) à obtenir. Nous proposons donc, dans ce travail, d'utiliser un lexique issu d'une langue proche et mieux dotée, pour amorcer la construction d'un système G2P pour une nouvelle langue.

En particulier, nous proposons d'utiliser les prononciations initiales produites, sur un lexique en langue iban, par un système G2P malais puis de faire postéditer (corriger) ces prononciations par des locuteurs natifs iban. Nous pensons cette approche raisonnable car malais et iban sont étroitement liés et tous deux appartiennent à un même groupe de langues d'après (Lewis *et al.*, 2013) (Adelaar, 2005) . Dans un premier temps, nous analysons la "distance" entre les prononciations générées par notre système G2P en malais et les corrections de ces prononciations par un locuteur natif iban. Dans (Heeringa et de Wet, 2008), une étude similaire est menée pour mesurer les distances entre l'afrikaans et le néerlandais, l'afrikaans et le frison, ainsi que l'afrikaans et l'allemand. Les auteurs concluent que le néerlandais présente des prononciations plus proches de l'afrikaans que les deux autres langues. A partir des post-editions de prononciations obtenues sur un lexique initial en iban, nous construisons un système G2P empirique pour l'iban afin de phonétiser d'autres entrées. Les deux systèmes G2P malais et G2P iban, ainsi qu'une méthode hybride (fondée sur l'observation que de nombreux mots du malais sont utilisés en iban), sont tout d'abord évalués. Les lexiques phonétisés obtenus sont ensuite intégrés dans le premier système de RAP jamais développé pour l'iban.

Cet article est organisé comme suit : la section 2 décrit brièvement la phonologie de la langue iban. La section 3 présente les ressources collectées pour cette langue (pour la RAP) ainsi que notre stratégie de construction d'un système G2P. La section 4 présente les différents modèles acoustiques construits pour l'iban à partir des 3 différents dictionnaires de prononciation obtenus et les performances de reconnaissance automatique de la parole associées. La partie 5 analyse plus en détail les sorties et les erreurs des différents systèmes de RAP évalués et enfin, la section 6 conclut et donne quelques perspectives à ce travail.

Une partie de cet article, consacrée à l'obtention d'un système G2P pour l'iban a déjà été publiée en anglais, en Octobre 2013, pour le *4th Workshop on South and Southeast Asian Natural Language Processing* qui s'est tenu en même temps que la conférence IJCNLP 2013. Cependant, contrairement à la présente soumission, cet article ne contenait encore aucune expérience de RAP.

2 La langue iban

L'iban[1] est une langue austronésienne (branche Malayo-Sumbawan) parlée par plus de 600 000 locuteurs sur l'île de Bornéo, principalement en malaisie (État de Sarawak), ainsi qu'au Brunei et en Indonésie. La langue appartient au même groupe que le malais (groupe *Malayic* - voir (Lewis *et al.*, 2013) et (Dryer et Haspelmath, 2013) . En plus de l'apprentissage du malais, langue officielle, l'iban est enseigné dans les écoles primaire et secondaire (mais non obligatoire) depuis le début des années 90. Dans plusieurs universités, notamment à l'université de Sarawak[2], des cours de langue iban débutants sont proposés aux étudiants de premier cycle. L'iban est influencé par le malais en termes de phonologie, de morphologie et de syntaxe (Education-Department, 2007). Selon (Education-Department, 2007), beaucoup de mots sont communs à ces deux langues (mêmes formes de surface) et il y a de nombreux emprunts au malais. Quelques exemples sont répertoriés dans le tableau 1 avec les prononciations associées.

TABLE 1 – *Exemples de mots communs malais/iban avec les prononciations associées*

No.	Mots	iban	malais
1	ke	/kə/	/kə/
2	nya	/ɲaʔ/	/ɲə/ or / ɲa/
3	iya	/ija/	/ija/
4	bilik	/biliəʔ/	/bileʔ/
5	dua	/duwa/	/duwə/ or /duwa/
6	sida	/sidaʔ/	/sida/
7	puluh	/puluəh/	/puloh/
8	raban	/raban/	/raban/
9	lalu	/lalu/	/lalu/
10	orang	/uraŋ/	/oraŋ/

En 1981, (Omar, 1981) publie une description complète de cette langue. Selon l'auteur, la langue comprend 19 consonnes (/p/, /b/ /t/, /d/, /k/, /g/, /ʔ/, /ʧ/, /dʒ/ , /s/, /h/, /m/, /n/, /ɲ/, /ŋ/, /l/, /r/, /j/, /w/), 6 voyelles (/i/, /e/, /a/, /o/, /u/, / ə/) et 11 séquences de voyelles[3] (/ui/, /ia/, /ea/, /ua/, /oa/, /iu/, /iə/, /uə/, /oə/, /ai/, /au/). Cette liste n'inclut pas les consonnes empruntées au malais telles que /f/, /v/, /θ/, /ð/, /z/, /ʃ/, /x/, /ɣ/. Comme le montre la Table 1, iban et malais utilisent le système d'écriture Latin ; par ailleurs, ce ne sont pas des langues tonales. En dehors des séquences de voyelles, bien moins nombreuses en malais, une autre différence entre les deux langues est la présence plus fréquente du son /ʔ/ à la fin de mots iban se terminant par une voyelle.

3 Ressources pour l'iban

3.1 Corpus textuel

Des données textuelles de type news ont été collectées sur un site web d'information[4] qui publie quelques articles en iban chaque jour. Nous avons récupéré ces articles pour la période 2009-2012 et rassemblé un total de 7000 articles dans divers domaines (general, sport et divertissement).

1. iban language code : [iba] (ISO 639-3)
2. http ://www.unimas.my
3. nommées *vowel clusters* par (Omar, 1981)
4. www.theborneopost.com/news/utusan-borneo/berita-Iban/

Après extraction, le texte a été nettoyé et normalisé par les étapes suivantes : (1) enlever les tags HTML, (2) convertir les dates et les nombres en mots (e.g : 1982 à *sembilan belas lapan puluh dua*), (3) convertir les abbréviations (e.g : Dr. à *Doktor*, Prof. à *Profesor*, Kpt. à *Kapten*), (4) segmenter en phrases, (5) mettre en minuscules (6) enlever les marques de ponctuation à part les tirets. Au final, un corpus de 2.08M mots est disponible. C'est peu pour la modélisation statistique du langage, mais c'est à ce jour le seul corpus textuel de plus d'un million de mots disponible pour cette langue.

Grâce à ces données de texte, nous avons construit un modèle de langage à base de trigrammes avec la boîte à outils SRILM (Stolcke, 2002). La perplexité du modèle, mesurée sur les transcriptions du corpus oral collecté (voir paragraphe suivant) est de 162 et le taux de mots hors vocabulaire (%MHV) est de 2,3 %.

3.2 Corpus oral transcrit

Nous disposons de 8 heures transcrites de parole échantillonnée à 16kHz. Les données ont été fournies par la RTM (*Radio Televisyen Malaysia*), une station de radio et de télévision locale, à travers l'une de ses chaînes radio appelée WaiFM. La chaîne diffuse cinq à dix minutes de nouvelles quotidiennes en langue iban. La table 2 et la table 3 donnent plus de détails sur le corpus oral et sa répartition entre apprentissage et test.

TABLE 2 – iban : corpus oral

Genre	Locuteurs	Phrases	Tokens	Durée (mins)
Femme	14	1,382	36,194	222
Homme	9	1,750	44,408	257

TABLE 3 – Répartition apprentissage / test pour nos expériences

Set	Locuteurs	Genre (H :F)	Phrases	(mins)
Test	6	2 :4	473	71
Apprentissage	17	7 :10	2659	408

Les données de parole ont été transcrites par 8 locuteurs iban natifs (dont 7 femmes). Au cours d'un atelier préliminaire tenu à l'Université de Sarawak, les transcripteurs ont reçu une session de formation sur Transcriber ((Barras *et al.*, 2000), (Barras *et al.*, 1998)). Au total, on dispose de 3132 phrases transcrites, prononcées par 23 locuteurs différents. Un sous ensemble de 473 phrases, représentant environ 1h de corpus, est choisi pour l'évaluation.

3.3 Dictionnaire de prononciation

3.3.1 Construction du système G2P pour l'iban

Notre corpus textuel iban contient 37k mots différents. Parmi ces mots iban, on trouve aussi du malais (23%) et de l'anglais (19%). Ces statistiques sont obtenues en comparant notre liste de mots iban avec un dictionnaire malais disponible et issu de (Tan *et al.*, 2009) et avec le dictionnaire de CMU en anglais.

Le lexique malais initial de (Tan *et al.*, 2009) nous permet d'entrainer un système G2P pour le malais en utilisant Phonetisaurus, un outil open source fondé sur des transducteurs à états finis ((Novak, 2012), (Novak *et al.*, 2011)) . Un sous-ensemble de 68k est prélevé du dictionnaire initial de 76k tandis que 8k mots sont conservés pour l'évaluation du système G2P construit. Le taux d'erreur de phonèmes (PER) est de 6,20 % tandis que le taux d'erreur de mots (WER) est de 24,98 % (voir (Juan et Besacier, 2013) pour plus de détails).

Par la suite, avec une méthodologie similaire à celle de (Heeringa et de Wet, 2008), nous sélectionnons les 100 mots les plus fréquents de notre corpus textuel et appartenant à la fois à notre lexique iban (37k) et au lexique malais de (Tan *et al.*, 2009) (76k prononciations disponibles). Nous indiquerons à l'avenir *iban-malais* pour faire référence à ces mots communs aux deux lexiques iban et malais. Le système G2P malais est utilisé pour phonétiser nos 100 mots sélectionnés et les sorties du système G2P malais sont ensuite corrigées par un locuteur natif iban [5]. La comparaison entre les sorties corrigées indique un taux d'erreur de mots de 47 % (WER) ce qui signifie que 53 prononciations sur les 100 ont été considérées comme parfaites par l'annotatrice. Ce système G2P malais peut donc être considéré comme un point de départ acceptable pour produire des prononciations pour la langue iban.

En conséquence, le système G2P malais est alors utilisé pour phonétiser un sous-ensemble de 1k du lexique iban initial. Parmi les 1k mots, 500 sont des mots iban non existant en malais, tandis que 500 autres sont de la catégorie *iban-malais* définie précédemment. Les sorties du système G2P malais sont à nouveau corrigées et les 1k mots correctement phonétisés en iban sont ensuite utilisés pour construire un premier système G2P en iban, toujours avec Phonetisaurus. Notre évaluation, sur un autre sous-ensemble de 1k mots, est présentée dans la table 4 (issue de (Juan et Besacier, 2013)). Les résultats montrent que le nouveau système G2P obtenu pour l'iban est plus performant pour les mots spécifiques à cette langue (notés avec l'indice I dans la table) tandis que le système G2P malais phonétise mieux les mots communs aux deux langues (notés avec l'indice IM dans la table).

TABLE 4 – Phonétisation de l'iban : performances des systèmes G2P malais et G2P iban

Phonetiseur	Corpus	PER (%)	WER (%)	Post-edition (mins)
G2P malais	500_{IM}	6.52	27.2	30
	500_{I}	15.8	56.0	42
G2P iban	500_{IM}	13.6	44.2	45
	500_{I}	8.2	31.8	32

Note : IM pour mots malais-iban et I pour mots "pur iban". (Juan et Besacier, 2013)

3.3.2 Obtention d'un lexique iban complet

Le lexique complet iban (37k) est phonétisé avec plusieurs systèmes G2P. En plus d'un premier lexique iban phonétisé avec le système G2P malais (S1) et d'un second lexique iban phonétisé avec le système G2P iban (S2), nous considérons aussi un système G2P hybride (S3) qui repose sur la stratégie suivante : le système G2P malais phonétise les mots de type *iban-malais* tandis que le système G2P iban phonétise les autres mots ("pur" iban). Par ailleurs, les mots d'origine anglaise sont phonétisés par le système G2P malais (comme montré dans (Tan *et al.*, 2009) car le lexique ayant servi à l'apprentissage de ce système comprend beaucoup de mots anglais). Au final, nous disposons d'un vocabulaire iban de 37k mots phonétisés (1K manuellement et les

5. la première auteure de cet article

autres 36k automatiquement). Une dernière évaluation réalisée sur 2000 nouveaux mots indique un taux d'erreur PER de 8.1% et un WER de 29.4% obtenu avec le système hybride S3. Dans la section suivante, ces trois systèmes G2P (S1, S2 et S3) sont évalués dans le cadre d'un système de reconnaissance automatique de la parole.

4 Systèmes de reconnaissance automatique de la parole (RAP)

4.1 Systèmes Kaldi réalisés

Les systèmes sont développés pour l'iban à partir de l'outil open-source Kaldi (Povey *et al.*, 2011). Les modèles acoustiques sont appris sur la partie apprentissage du corpus décrit au paragraphe 3.2 avec les 3 dictionnaires de prononciations issus des systèmes G2P S1 (malais), S2 (iban) et S3 (Hybride). Nous reportons les performances obtenues avec les différentes étapes de l'apprentissage. Les paramètres extraits sont 13 MFCCs pour les premiers modèles de type monophone et triphone (4,200 états et 40,000 Gaussiennes). L'enrichissement des paramètres par les coefficients delta et delta-delta, ainsi que plusieurs transformations telles que LDA et MLLT (Gopinath, 1998) sont aussi appliquées et évaluées. Enfin, un apprentissage adapté au locuteur (*Speaker Adaptive Training* - SAT) est réalisé et combiné avec une adaptation dans l'espace des paramètres de type fMLLR (Gales, 1998). Le décodage est réalisé avec plusieurs facteurs de poids pour le modèle de langage (16 au total) et nous reportons ici les meilleurs résultats obtenus pour chaque configuration.

4.2 Performances

Les résultats sont résumés dans la Table 5. Les performances augmentent progressivement au fur et à mesure des étapes. Il est suprenant de constater que la différence entre les trois systèmes, qui utilisent des stratégies de phonétisation relativement différentes, est en fait très faible. Les meilleures performances sont bien obtenues avec la méthode de phonétisation la plus pointue (S3 - Hybride - 20.6% WER) mais la différence est très faible avec le système fondé sur le phonétiseur malais (S2 - malais - 20.82% WER). Ceci montre que l'utilisation d'un phonétiseur déjà disponible dans une langue proche est une option tout à fait viable pour amorcer le développement d'un système de RAP dans une nouvelle langue très peu dotée.

TABLE 5 – RAP de la langue iban - Performances (WER%) pour différents systèmes G2P

Etapes	Lexique		
	malais G2P	iban G2P	Hybride G2P
Monophone	42.17	41.79	41.97
Triphone	36	36.44	36.11
Triphone + Δ + Δ	36.47	36.98	36.77
+ MLLT + LDA	27.24	27.71	26.80
+ SAT	25.86	27.02	25.82
+ fMLLR	20.82	21.90	20.60

4.3 Combinaison de systèmes

Afin de voir si les systèmes construits à partir de plusieurs lexiques sont complémentaires, nous combinons les treillis issus des 3 meilleurs systèmes (dernière ligne de la table 5) et utilisons la technique de décodage par minimisation du risque de Bayes (MBR (Goel *et al.*, 2003)). Les

résultats sont présentés dans la table 6 pour la combinaison entre 2 ou 3 systèmes. On observe une amélioration des résultats de 1.38 % en absolu dans le meilleur des cas (combinaison de trois systèmes). Au final, et compte tenu de la taille limitée de notre corpus d'apprentissage, nous considérons que ce premier système de RAP de la langue iban est assez performant. Ses erreurs sont cependant analysées plus en détail dans le paragraphe suivant.

TABLE 6 – Combinaisons de systèmes de RAP en utilisant différentes stratégies de phonétisation

Combinaison	%WER
Hybride G2P + iban G2P	19.83
malais G2P + iban G2P	19.76
Hybride G2P + malais G2P	19.55
Hybride G2P + malais G2P + iban G2P	19.22

5 Analyse des résultats obtenus

5.1 Faible corrélation entre performances de RAP et de G2P

La figure 1 présente les performances de reconnaissance automatique de la parole (WER en ordonnées) en fonction des performances du système de phonétisation G2P (à nouveau, WER en abcisse). Il est intéressant de constater la relative faible corrélation entre les performances des systèmes. En d'autres termes, les systèmes de phonétisation (G2P) ne semblent avoir que très peu d'influence sur les performances du système de transcription automatique. La bonne nouvelle est que l'utilisation d'une langue similaire dotée de ressources (comme le malais) semble être un bon point de départ pour générer les prononciations et construire un système de reconnaissance pour une langue très peu dotée (comme l'iban).

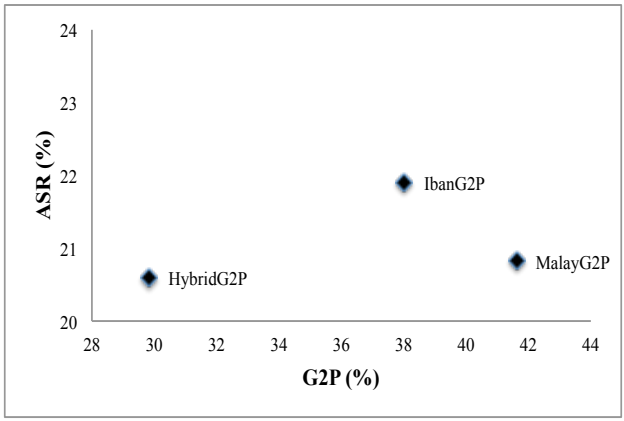

FIGURE 1 – Performance de RAP (ASR) en fonction de la performance de phonétisation (G2P) pour l'iban

5.2 Analyse des erreurs

Le Tableau 7 présente les principales confusions (top 10) pour les systèmes fondés sur un G2P Hybride, malais, iban ; ainsi que sur un système combiné. Il apparaît que les principales erreurs sont dues à des problèmes de normalisation othographique et de découpage morphologique. Par

TABLE 7 – Principales confusions (top 10) pour les systèmes fondés sur un G2P Hybride, malais, iban ; ainsi que sur un système combiné

Hybride	malais	iban	Combiné (H+M+I)
rakyat => rayat	rakyat => rayat	rakyat => rayat	rakyat => rayat
ka => ke	ka => ke	ari => hari	ka => ke
ti => ke	ari => hari	ka => ke	ari => hari
ari => hari	serta => sereta	serta => sereta	ti => ke
urang => orang	ti => ke	ti => ke	serta => sereta
serta => sereta	urang => orang	urang => orang	urang => orang
mohamad => mohd	datuk => dato	ke => ka	ke => ka
ka => madahka	ka =>madahka	mohamad => mohd	mohamad => mohd
ke => bejalaika	ke => bejalaika	agensi => ijinsi	agensi => ijinsi
antara => entara	mohamad => mohd	ka => madahka	ka => madahka

exemple, le mot *rakyat* (people) est un mot malais reconnu par le système comme *rayat* qui signifie la même chose en iban. C'est donc une erreur dans la référence qui pénalise l'évaluation du système de RAP. Une raison possible est que les transcripteurs ont été influencés par l'orthographie du malais. Les autres exemples sont *urang* et *orang* (person), *serta* et *sereta*, *mohamad* et *mohd*, *penerbai* et *penerebai* (airline), *agensi* et *ijinsi* (agency). Des confusions liées à un problème de découpage morphologique sont par exemple *ka* et *madahka*, *ka* et *bejalaika*, *ka* et *ngambika*, *waifm* et *fm*. *ka* est un suffixe utilisé dans les verbes transitifs et a la même fonction que *kan* en malais. Apparemment, ce suffixe est souvent séparé de sa racine dans nos textes iban et dans nos transcriptions de parole. Enfin, les différentes colonnes (correspondant à différentes comparaisons entre systèmes) ne font pas apparaître des erreurs fondamentalement différentes d'un système à l'autre. Encore une fois, ceci semble indiquer que le dictionnaire de prononciation a finalement peu d'influence sur les principales erreurs produites par le système de RAP.

6 Conclusions et perspectives

Cet article a décrit notre collecte de ressources pour la langue iban, dans l'objectif de construire un système de reconnaissance automatique de la parole pour cette langue. Nous nous sommes plus particulièrement focalisés sur une méthodologie d'amorçage du lexique phonétisé à partir d'une langue proche (le malais). Les performances des premiers systèmes de reconnaissance automatique de la parole construits pour l'iban sont encourageantes et montrent aussi que la qualité du dictionnaire de prononciation a une influence limitée sur la qualité de transcription. Ce résultat peut apparaître comme positif car il suggère une méthode générique pour amorcer un système G2P en utilisant une langue proche. Une première analyse des erreurs fait ressortir des problèmes bien connus pour les langues peu dotées : problèmes de normalisation de l'orthographe, erreurs liées à la morphologie (séparation ou non des affixes de la racine), etc. C'est notamment sur ces points que se porteront nos efforts dans le futur. Nous venons, par exemple, de lancer une campagne d'annotation collaborative de données iban où il est demandé aux annotateurs d'indiquer d'éventuels problèmes de normalisation et de choisir, à partir d'un couple de mots similaires, la bonne orthographie. Cette campagne vient d'être lancée sur le groupe Facebook *Iban People*.

Références

ADELAAR, A. (2005). The austronesian languages of asia and madagascar. *In The Austronesian Languages of Asia and Madagascar : A Historical Perspective*, pages 1–42, London, UK. Routledge Language Family Series.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. *In Proc. First International Conference on Language Resources and Evaluation (LREC)*, pages 1371–1376.

BARRAS, C., GEOFFROIS, E., WU, Z. et LIBERMAN, M. (2000). Transcriber :development and use of a tool for assisting speech corpora production. *In Proc. Speech Communication special issue on Speech Annotation and Corpus Tools*, volume 33. available at : trans.sourceforge.net/en/publi.php.

BESACIER, L., BARNARD, E., KARPOV, A. et SCHULTZ, T. (2014). Automatic speech recognition for under-resourced languages : A survey. *Speech Communication Journal*, 56:85–100.

DAVEL, M. et MARTIROSIAN, O. (2009). Pronunciation dictionary development in resource-scarce envionments. *In In Proc. INTERSPEECH*, pages 2851–2854.

DRYER, M. et HASPELMATH, M. (2013). The world atlas of language structures online. leipzig : Max planck institute for evolutionary anthropology. available at : http ://wals.info.

EDUCATION-DEPARTMENT, S. (2007). *Sistem Jaku Iban di Sekula*. Sarawak, Malaysia, 1st édition.

GALES, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language*, 12:75–98.

GOEL, V., KUMAR, S. et BYRNE, W. (2003). Segmental minimum bayes-risk decoding for automatic speech recognition. *In IEEE Transactions on Speech and Audio Processing*.

GOPINATH, R. A. (1998). Maximum likelihood modeling with gaussian distributions for classification. *In Proceedings of ICASSP*, pages 661–664.

HEERINGA, W. et de WET, F. (2008). The origin of afrikaans pronunciation : a comparison to west germanic languages and dutch dialects. *In Proc. Conference of the Pattern Recognition Association of South Africa*, pages 159–164.

JUAN, S. S. et BESACIER, L. (2013). Fast bootstrapping of grapheme to phoneme system for under-resourced languages - application to the iban language. *In Proc. 4th Workshop on South and Southeast Asian Natural Language Processing 2013*, Nagoya, Japan.

LEWIS, M. P., SIMONS, G. F. et FENNIG, C. D. (2013). Ethnologue : Languages of the world, sil international. available at : http ://www.ethnologue.com.

MASKEY, S. R., BLACK, A. W. et TOMOKIYO, L. M. (2004). Bootstrapping phonetic lexicons for language. *In Proc. INTERSPEECH*, pages 69–72.

NOVAK, J. R. (2012). Phonetisaurus : A wfst-driven phoneticizer. available at : https ://code.google.com/p/phonetisaurus.

NOVAK, J. R., MINEMATSU, N. et HIROSE, K. (2011). Evaluations of an open source wfst-based phoneticezer. PDF, General Talk No. 452, The Institute of Electronics, Information and Communication Engineers.

OMAR, A. H. (1981). *Phonology*. Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Schwarz, O. G. N. G. M. H. P., Silovsky, J., Stemmer, G. et Vesely, K. (2011). The kaldi speech recognition toolkit. *In* Society, I. S. P., éditeur : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, volume IEEE Catalog No. : CFP11SRW-USB.

Stolcke, A. (2002). Srilm - an extensible language modeling toolkit. *In Proc. of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Tan, T.-P., Li, H., Tang, E. K., Xiao, X. et Chng, E. S. (2009). Mass : a malay language lvsr corpus resource. *In Proc. Oriental COCOSDA International Conference 2009*, pages 26–30.