Expériences de transcription automatique d'une langue rare

Thomas Pellegrini and Lori Lamel

LIMSI-CNRS, BP133
91403 Orsay cedex, FRANCE
{thomas.pellegrini, lamel}@limsi.fr

ABSTRACT

This work investigates automatic transcription of rare languages, where rare means that there are limited resources available in electronic form. In particular, some experiments on word decompounding for Amharic, as a means of compensating for the lack of textual data are described. A corpus-based decompounding algorithm has been applied to a 4.6M word corpus. Compounding in Amharic was found to result from the addition of prefixes and suffixes. Using seven frequent affixes reduces the out of vocabulary rate from 7.0% to 4.8% and total number of lexemes from 133k to 119k. Preliminary attempts at recombining the morphemes into words results in a slight decrease in word error rate relative to that obtained with a full word representation.

1. INTRODUCTION

Les termes utilisés dans la littérature pour désigner les langues rares sont variés, que ce soit en français : langues rares, langues peu-dotées, langues peu-informatisées (langues pi), langues minoritaires, ou que ce soit en anglais : under-represented languages, under-resourced languages, less widely available languages ou encore minority languages. Cette diversité de vocabulaire reflète le caractère subjectif de la notion de rareté pour les langues, du point de vue du traitement automatique. Dans le rapport final du projet européen INTERA [5] qui visait la création de nouveaux corpus de ressources multilingues pour des langues européennes, les langues visées furent des langues "moins accessibles du point de vue numérique". Cette définition relève d'une comparaison implicite avec les langues dominantes, et parait plus précise que le simple adjectif "rare" employé dans un souci de concision dans cet article. L'amharique, langue officielle de l'Ethiopie, fait l'objet de recherches récentes en traitement automatique telles que le classement thématique de textes [2], les outils d'analyse morphologique [3] et la transcription automatique [6],[9]. Pour cette langue, l'audio n'est pas un facteur limitant puisque de nombreuses radios diffusent quotidiennement leurs émissions sur Internet [1]. En revanche il s'agit d'une langue rare en raison de la faible quantité de textes disponibles sous forme électronique. Le corpus de textes utilisé dans cette étude ne dépasse pas 5 millions de mots, chiffre à comparer au milliard de mots utilisés pour les systèmes de transcription d'anglais américain et du français du LIMSI.

Dans un premier temps, les propriétés lexicales du corpus de textes et l'élaboration d'un lexique de prononciation seront décrites, les résultats d'un système standard de transcription automatique seront présentés. Dans un deuxième temps les premiers résultats d'expériences de décomposition des mots (par séparation d'affixes) seront discutés.

[1] Deutsche Welle et Radio Medhin par exemple

2. PRÉSENTATION DE L'AMHARIQUE

L'amharique est parlé par environ 14 millions de locuteurs. Bien que faisant partie des langues sémitiques comme l'arabe et l'hébreu, l'écriture se fait de gauche à droite avec un syllabaire spécifique dérivé de la langue classique éthiopienne, le ge'ez. L'amharique possède 34 symboles de base dont 85% représentent une séquence CV (C pour consonne, V pour voyelle), les autres symboles représentent une séquence CwV où w est une semi-consonne. Un dernier symbole représente le son complexe /ts/. Cette langue possède sept voyelles au total, schwa inclus, appelées les sept ordres : /ɛ/, /u/, /i/, /a/, /e/, /ə/ et /o/.

En ce qui concerne l'écrit, des problèmes de normalisation de l'orthographe amharique sont exposés dans [10], où trois niveaux de langue sont distingués :

- l'amharique canonique, réservé aux érudits (une orthographe unique par mot),
- l'amharique commun, celui des journaux, de la littérature (formes homophones),
- l'amharique quotidien, pas de jugement porté sur l'orthographe des mots.

L'orthographe des mots amhariques utilisés au quotidien est très libre, le nombre de formes écrites différentes pour un même mot peut être très grand. Des exemples d'orthographe ambigue rencontrés seront donnés ci-après.

3. LES RESSOURCES AUDIO ET TEXTES

Pour élaborer le système de transcription et réaliser les études décrites dans cet article, deux types de données ont été utilisés : un corpus d'émissions de radio transcrites et un corpus de textes issus de sites Web de journaux en ligne. Le corpus audio contient 37 heures d'émissions de type journal provenant de 2 sources : radio Deutsche Welle (25h) et radio Medhin (12h) enregistrées de janvier 2003 à février 2004. Ces données ont été transcrites manuellement par des locuteurs ethiopiens. Pour tester et développer le système de transcription automatique, deux heures de données audio transcrites ont été sélectionnées au sein du corpus. Il s'agit des fichiers audio parmi les plus récents du corpus, les thèmes abordés dans ces données pouvant être nouveaux par rapport à ceux des données d'apprentissage. Le tableau 1 résume les caractéristiques des données

audio : le nombre d'heures par source, le nombre de locuteurs et le nombre de mots pour le corpus d'apprentissage et pour le corpus de développement. Nous avons retenu un plus grand nombre d'heures provenant de Deutsche Welle que de radio Medhin, car les émissions de cette radio ont une plus grande diversité de locuteurs.

TAB. 1: Nombre d'heures, de locuteurs et de mots pour les deux sources audio

source	app	dev
Deutsche welle	24h06	1h20
radio Medhin	11h08	0h37
# locuteurs	200	15
# mots	232.6k	14.1k

Les données textuelles autres que les transcriptions manuelles proviennent de 3 sources : Ethiozena (archives de 1988 a 1996 et textes récents), Deutsche Welle (textes récents) et Ethiopian Reporter (textes récents). Au total pour ces trois sources nous disposons de 4.6 millions de mots. Les textes des transcriptions du corpus audio sont également utilisées et totalisent 246.7k mots.

4. PROPRIÉTÉS LEXICALES

A l'instar de l'arabe et de l'hébreu un grand nombre de mots, en particulier les verbes, se forment à partir de racines de trois lettres (racines tri-consonantiques) auxquelles s'ajoutent des schèmes (les voyelles) qui précisent le sens des mots ainsi formés [3]. Viennent s'ajouter des marques (de conjugaison, de pronoms personnels, de pluriel, etc…) qui donnent de nombreuses formes dérivées à partir d'une même racine.

La figure 1 montre le nombre de mots distincts en fonction de leur taille en nombre de phonèmes pour les transcriptions manuelles (courbe en trait plein) et pour les textes Web (courbe en pointillé). La taille de mots la plus fréquente est relativement grande puisqu'elle est de dix phonèmes soit cinq syllabes (cinq symboles amhariques), ce qui pourrait justifier la démarche de chercher des affixes pour décomposer les mots longs.

La figure 2 donne la taille moyenne des mots en fonction du rang de fréquence des mots distincts des textes Web. La courbe montre que les mots les plus fréquents sont les plus courts avec une taille entre deux et trois syllabes, ce qui est un comportement tout à fait naturel, observé pour une majorité de langues. Néanmoins le corpus de textes étant très petit, le nombre de mots distincts peu fréquents est grand. Pour un rang de fréquence compris entre 1 et 500, il y a un seul mot par rang. A partir de 500, ce nombre augmente et atteint, pour les mots les moins fréquents (1300^e rang, 3 occurences dans le corpus), une valeur supérieure à 26k mots. Pour les textes Web, seuls les mots apparaissant au moins trois fois sont gardés pour se débarrasser le plus possible des mots "parasites" (chiffres accolés aux mots, caractères non-identifiés, etc…)

Les dix premiers rangs de fréquence couvrent un peu plus de 5% des 4.6M de mots des textes Web (dix mots distincts totalisant 245.8k occurences). Les dix derniers rangs de fréquence couvrent 10% (80k mots distincts totalisant 423.9k occurences). Les expériences de décomposition sur les mots longs (moins fréquents mais nombreux) peuvent donc s'avérer intéressantes pour les performances du système de transcription, à condition que les racines des mots dont les affixes ont été séparés soient des mots déjà présents avant décomposition.

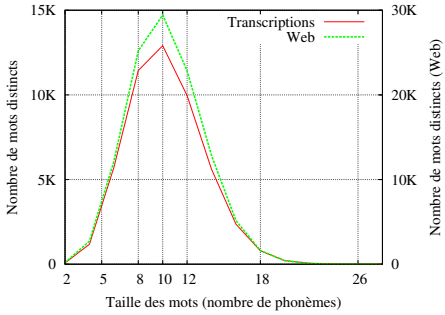

FIG. 1: Distribution des mots distincts des transcriptions et des textes Web en fonction de leur taille (en nombre de phonèmes)

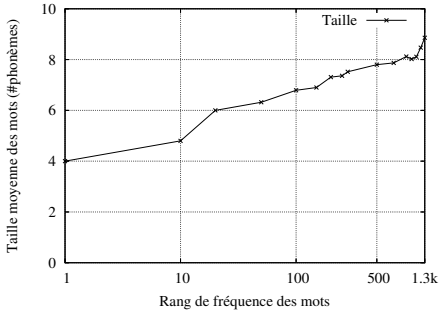

FIG. 2: Taille moyenne des mots distincts des textes Web (112k mots) en fonction de leur rang de fréquence

Le taux de mots hors vocabulaire (MHV) influence grandement les performances des systèmes de transcription (1 mot hors vocabulaire génère entre 1.5 et 2 erreurs). Les systèmes actuels travaillent avec des lexiques fixes qui doivent optimiser la couverture lexicale. Avec un lexique de 133k mots réalisé avec tous les mots des transcriptions et les mots des textes Web apparaissant au moins trois fois, le taux de MHV atteint 7.0% sur le corpus de développement de 14.1k mots. A titre de comparaison, le taux de MHV pour le système de transcription de l'anglais américain est d'environ 0.5% avec un lexique de 65k mots [4]. Tenter de le diminuer en décomposant les mots longs est une autre motivation importante pour cette étude.

5. LE LEXIQUE DE PRONONCIATIONS

Le jeu de phonèmes utilisé dans ces expériences comporte 33 phones (voyelles comprises) avec trois phones supplémentaires (pour les silences, les respirations et les hésitations, annotés dans les transcriptions). Les suites Cw sont modélisées avec deux unités distinctes, une pour C, une autre pour w. Si aux symboles amhariques qui codent les suites Cw avaient été associés un unique caractère de translitération, le nombre de phones aurait été plus grand. Dans le système de reconnaissance décrit dans [9] par exemple, le jeu de phones en compte 38 soit 5 consonnes de plus.

Les 240 symboles amhariques ont été translitérés en utilisant le jeu de phones mentionné ci-dessus. Voici un exemple de phrase tirée d'une transcription manuelle, il s'agit d'une phrase d'introduction des émissions d'information de la radio Medhin (? pour le coup de glotte, x pour le schwa) :

የኢትዮጵያ መድህን ድምፅ ራዲዮ

jE ?itxjoPxja mEdxhxnx dxmxtsx radijo

Comme les caractères amhariques sont translitérés avec un jeu de caractères représentant les phonèmes eux-mêmes, un premier dictionnaire de prononciation constitué de la simple liste des mots a été utilisé pour réaliser les premiers alignements. Ces alignements ont permis de voir que les schwas ne sont pas toujours prononcés et un deuxième dictionnaire avec tous les schwas optionnels a été utilisé. Dans l'article [8], des expériences de détection des variantes de prononciation des voyelles amhariques à l'aide d'alignements sont décrites. Des alignements de syllabes sont réalisés avec un lexique permettant la substitution et l'élision de toutes les voyelles. Globalement, plus de 60% des schwas sont éliminés. Le lexique utilisé dans cette étude est un lexique de 133k mots avec uniquement les schwas optionnels. Le tableau 2 donne un exemple de trois mots de ce lexique, avec leur rang de fréquence et leur nombre d'occurences au sein des transcriptions manuelles. Le schwa est mis entre accolades pour signifier son caractère optionnel dans les formes phonémiques du lexique. Sur les 3k occurences de "nEwx", 2.5k ont été alignées avec la prononciation sans schwa.

TAB. 2: Exemples de mots du lexique avec leur prononciation, rang de fréquence et nombre d'occurences

lexème	forme phonémique	rang	#occ.
nEwx	nEw{x}	1	3044
mEto	mEto	7	803
jEdimokxrasi	jEdimok{x}rasi	236	47

6. RÉSULTATS AVEC UN SYSTÈME STANDARD

Un système de transcription standard basé sur [4], en deux passes avec une adaptation non-supervisée des modèles acoustiques après la première passe a été construit avec la boîte à outils STK du LIMSI. Le modèle de langage est issu de l'interpolation de deux modèles avec lissage de Kneser-Ney, le premier est un modèle appris uniquement sur les transcriptions et le second sur les textes Web. La perplexité mesurée sur le corpus de développement est relativement élevée : $px = 372$ avec un taux de MHV de 7.0%. Les modèles acoustiques utilisés sont des triphones à états liés (avec 32 gaussiennes par état, 3 états par modèle) dépendants du contexte et de la position intermot, i.e. différents modèles sont utilisés pour des phones à l'intérieur des mots et pour des phones en frontière de mots. Au total, 10.7k contextes sont modélisés par 8.5k états. Les contextes peu observés sont regroupés en regardant en premier lieu s'ils ont un contexte droit en commun (co-articulation régressive), sinon un contexte gauche en commun (co-articulation progressive) et enfin ceux pour lesquels aucun contexte en commun n'a été trouvé sont regroupés en modèles indépendants du contexte.

Le taux d'erreur sur les mots évalué à l'aide de l'outil sclite de NIST sur le corpus de développement est de 25.9%. Le tableau 3 précise les pourcentages des trois types d'erreur (E pour élision, S pour substitution et I pour insertion). L'équilibre entre les taux d'élisions et d'insertion est obtenu par une phase de réglage des paramètres du décodeur (le poids du modèle de langage, les pénalités sur le nombre de mots, le nombre de silences entre autres).

Le tableau 4 montre des exemples d'erreurs fréquentes en donnant la référence puis l'hypothèse erronnée avec les types d'erreur correspondants. Dans le premier cas le système a reconnu un mot composé plutôt que deux mots plus petits ce qui a donné une élision et une substitution. Le deuxième exemple est l'inverse du premier. Le système a opté pour deux mots plus petits plutôt qu'un seul mot ce qui a donné une insertion et une substitution. Enfin le troisième exemple montre un problème a priori de normalisation orthographique, les deux mots ne différant que d'une voyelle, néanmoins les deux formes sont présentes dans les textes.

TAB. 3: Répartition des erreurs. Types d'erreur : E élision, S substitution et I insertion

S	E	I	Total
20.9%	2.6%	2.3%	25.9%

TAB. 4: Exemples d'erreurs fréquentes. REF : référence, HYP : hypothèse. Types d'erreur : E élision, S substitution et I insertion

	exemple	type d'erreur
REF	bEjE KEnu	
HYP	bEjEKEnu	E S
REF	?xnxdEdxlx	
HYP	?xnxdE dxlx	I S
REF	kE ?**e**konomiwx	
HYP	kE ?**i**konomiwx	S

7. SÉPARATION D'AFFIXES : PREMIERS RÉSULTATS

La taille du lexique et le fort taux de MHV peuvent être diminués en séparant des affixes. Des expériences de décomposition de mots ont été réalisées pour des langues où le procédé de composition des mots est très important, pour l'allemand par exemple [1]. En amharique la composition se limite à des morphèmes grammaticaux de type articles possessifs, démonstratifs, pronoms, prépositions et postpositions.

7.1. Le choix des affixes

Détecter les affixes automatiquement a l'avantage de ne pas utiliser de connaissances linguistiques spécifiques à la langue cible et rend la méthode portable à d'autres langues facilement. L'algorithme de Harris [7] est un algorithme de détection des frontières de morphèmes indépendant de la langue, nécessitant simplement un corpus de mots de la langue cible. Il exploite le fait qu'un début de mot de k caractères a naturellement peu de caractères successeurs distincts possibles pour former des mots qui existent dans la langue traitée pour k suffisamment grand. Au rang k+1 ce nombre réduira davantage. Si ce nombre augmente subitement pour un début de mot de k caractères alors ce début de mot est un morphème candidat, pouvant se composer avec d'autres morphèmes commençant par des lettres distinctes variées. Ainsi l'algorithme compte le nombre de caractères successeurs distincts possibles pour tous les débuts de mots de taille k et propose des frontières de morphèmes pour ces mots lorsqu'un maxima local est trouvé. Il ne s'agit pas de réaliser une analyse morphologique mais de dégager quelques affixes potentiels les plus fréquents. Séparer ces affixes des mots du lexique permet de réduire le nombre de lexèmes et d'augmenter la représentation de certains n-grammes peu observés [1].

Une liste de sept affixes (cinq préfixes et deux suffixes) a été retenue pour les premières expériences rapportées ici.

Les sept affixes sont les plus fréquents parmi ceux détectés par l'algorithme. Les affixes sont séparés des mots dont la taille après séparation est d'au moins deux syllabes. Un signe "+" est accolé aux affixes pour pouvoir recombiner les mots par la suite. Le tableau suivant donne la liste des affixes :

TAB. 5: Préfixes et affixes retenus

préfixes (5)	suffixes (2)
?xnxdE+	+CEwx
?xnxda+	+mx
?xnxdi+	
?xnxdx+	
jE+	

Le nombre de mots du lexique est réduit de plus de 11%, de 133k à 119k mots. Le taux de MHV diminue de 7.0% à 4.8% soit une réduction absolue de 2.2%.

7.2. Résultats

De nouveaux modèles acoustiques ont été appris pour la représentation avec affixes séparés et un nouveau modèle de langage a été généré. Le sytème de reconnaissance est le même que celui qui a servi pour la représentation en mots entiers, seuls les modèles acoustiques, le modèle de langage et le lexique de prononciation diffèrent. Les résultats suivants ont été obtenus avec 10.6k modèles acoustiques (8.7k états).

Le tableau 6 donne les taux d'erreur pour la représentation avec les affixes séparés et après recombinaison des mots. Le taux obtenu avec affixes séparés est nettement inférieur car le nombre de mots est plus grand avec cette representation (17.3k mots), les affixes étant très bien reconnus. En recombinant les affixes (grâce au signe "+" accolé), le taux d'erreur augmente. Un gain absolu de 0.8% est néanmoins observé par rapport au taux d'erreur de 25.9% avec le système appris sur les mots entiers (appelé S_{mots} par la suite).

TAB. 6: Taux d'erreur sur les mots avant et après recombinaison des affixes

représentation	taux d'erreur
affixes séparés	21.6%
mots recomposés	25.1%

Le tableau 7 donne un exemple où le système avec affixes séparés a été meilleur. Les phrases de référence, en gras dans le tableau, ont respectivement trois mots pour la représentation en mots entiers (S_{mots}) et quatre mots pour la représentation avec affixes séparés ($S_{affixes}$). Le système S_{mots} n'a pas bien reconnu la phrase, la sortie obtenue ayant deux mots au lieu de trois. En revanche le système $S_{affixes}$ a correctement reconnu la phrase de référence. Le tableau donne, pour chaque système, la log-vraisemblance (log-v) pour la phrase correcte (phrase de référence) et pour la phrase erronée résultat du décodage par le système S_{mots}. Pour $S_{affixes}$, la phrase erronée est la phrase erronée de S_{mots} après séparation de l'affixe.

La vraisemblance, qui est la probabilité des suites de mots, est utilisée par le décodeur pour sélectionner la meilleure hypothèse. Les mots "lajx" et "jE+" sont parmi les mots les plus fréquents des textes avec affixes séparés, alors que le mot " ?iraKxlajx"est beaucoup moins fréquent, ce qui favorise la phrase correcte, obtenue par $S_{affixes}$, qui contient "lajx"et "jE+". Pour le système S_{mots}, la forte probabilité de l'unigramme "lajx" ne suffit pas à favoriser

TAB. 7: Exemple de phrase correctement reconnue par $S_{affixes}$ mais erronée pour S_{mots}, comparaison des log-vraisemblances. Les lignes S_{mots} et $S_{affixes}$ donnent les sorties respectives des systèmes

sytème	phrase			log-v
S_{mots}	?iraKxlajx		jESxgxgxrx	-9.5551
	?iraKx	**lajx**	**jESxgxgxrx**	-9.6559
$S_{affixes}$	**?iraKx**	**lajx**	**jE+ Sxgxgxrx**	-9.2613
	?iraKxlajx		jE+ Sxgxgxrx	-10.1367

la phrase de référence. Le mot "jESxgxgxrx" étant rare, la séparation de l'affixe "jE+" a été bénéfique.

8. PERSPECTIVES

Dans cet article, un système de transcription de l'amharique a été présenté et a servi de référence pour évaluer un deuxième système construit avec une représentation des mots différente, avec séparation d'affixes. La séparation d'un très petit nombre d'affixes a permis de réduire la taille du lexique de plus de 11%, le taux de MHV de 2.2% absolu et d'observer un gain de 0.8% absolu sur le taux d'erreur sur les mots. De nouvelles expériences avec un nombre d'affixes plus grand seraient très intéressantes à mener, la question de la sélection des affixes étant problématique (leur nombre, leur ressemblance phonémique). Le découpage des mots augmente la représentation de certains ngrammes peu observés et semble être une méthode prometteuse pour aborder le problème de la transcription automatique de langues rares.

RÉFÉRENCES

[1] M. Adda-Decker. A corpus-based decompounding algorithm for German lexical modeling in LVCSR. In *Proc. Eurospeech*, Geneva, 2003.

[2] S. Eyassu and B.Gamback. Classifying Amharic news texts uing self-organizing Maps. In *ACL05 Workshop on computational Approaches to Semitic Languages*, Ann Arbor, Michigan, 2005.

[3] S. Fissaha and J. Haller. Amharic verb lexicon in the context of machine translation. In *TALN 2003*, Batz-sur-Mer, 2003.

[4] J.L. Gauvain, L. Lamel, and G. Adda. The LIMSI Broadcast News Transcription System. *Speech Communication*, 37 :89–108, 2002.

[5] M. Gavrilidou, V. Giouli, E. Desipri, M. Monachini, and C. Soria. Report on the model of LRs production.

[6] B. Gamback H. Seid. A speaker independent continuous speech recognizer for Amharic. In *Proc. Interspeech*, Lisboa, 2005.

[7] Z. Harris. From Phoneme to Morpheme. In *Language 31*, pages 190–222, 1996.

[8] T. Pellegrini and L. Lamel. Experimental detection of vowel pronunciation variants in Amharic. In *Proc. LREC*, Genoa, 2006.

[9] W. Menzel S.T. Abate and B. Tafila. An Amharic Speech Corpus for Large Vocabulary Continuous Speech Recognition. In *Proc. Interspeech*, Lisboa, 2005.

[10] D. Yacob. Application of the Double Metaphone Algorithm to Amharic Orthography. In *International Conference of Ethiopian Studies XV*, 2003.