Développement de ressources en swahili pour un système de reconnaissance automatique de la parole

Hadrien Gelas[1, 2] Laurent Besacier[2] François Pellegrino[1]
(1) Laboratoire Dynamique Du Langage, CNRS - Université de Lyon, France
(2) Laboratoire Informatique de Grenoble, CNRS - Université Joseph Fourier Grenoble 1, France
{hadrien.gelas, francois.pellegrino}@univ-lyon2.fr, laurent.besacier@imag.fr

RÉSUMÉ

Cette contribution décrit notre travail sur la production de ressources en swahili pour un système de reconnaissance automatique de la parole (RAP). Le swahili est une langue bantu parlée dans une vaste région d'Afrique de l'Est. Nous introduisons en premier lieu le statut de la langue. Ensuite, nous reportons les différentes stratégies choisies pour développer un corpus de texte, un dictionnaire de prononciation et un corpus de parole pour cette langue peu dotée. Nous explorons des méthodologies telles que le crowdsourcing ou un processus de transcription collaboratif. De plus, nous tirons avantage de certaines caractéristiques linguistiques comme la morphologie riche de la langue ou la part de vocabulaire partagée avec l'anglais, afin d'améliorer les performances de notre système de RAP de référence dans une tâche de transcription de parole radiodiffusée.

ABSTRACT

Developments of Swahili resources for an automatic speech recognition system

This article describes our efforts to provide ASR resources for Swahili, a Bantu language spoken in a wide area of East Africa. We start with an introduction on the language situation. Then, we report the selected strategies to develop a text corpus, a pronunciation dictionary and a speech corpus for this under-resourced language. We explore methodologies as crowdsourcing or collaborative transcription process. Besides, we take advantage of some linguistic characteristics of the language such as rich morphology or shared vocabulary with English to improve performance of our baseline Swahili ASR system in a broadcast speech transcription task.

MOTS-CLÉS : Swahili, langues peu dotées, reconnaissance automatique de la parole, ressources numériques.

KEYWORDS: Swahili, under-resourced languages, automatic speech recognition, resources.

1 Introduction

Durant ces dernières décennies, entrainées par l'évolution permanente de l'informatique et la constante informatisation de nos sociétés, les technologies du langage et de la parole ont connu des progrès majeurs. Le déploiement de ces technologies pour des langues peu-dotées représente un challenge important. En effet, l'utilité et les différentes applications de ces outils dans les pays en voie de développement se sont montrées nombreuses : autant pour l'accès à l'information

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 633–640,
Grenoble, 4 au 8 juin 2012.

en Afrique Sub-Saharienne (Barnard *et al.*, 2010), pour l'agriculture en Inde rurale (Patel *et al.*, 2010) ou la santé au Pakistan (Kumar *et al.*, 2011). Cependant, les technologies du langage sont toujours pleinement confrontées au manque de ressources numériques et représentent ainsi peu la diversité des langues (plus de 6000). Comme beaucoup d'autres, les langues d'Afrique sont fortement touchées par cette lacune.

Le swahili est une importante langue véhiculaire d'Afrique de l'Est couvrant un large territoire de plus de huit pays (langue nationale au Kenya et en Tanzanie)(Polomé, 1967). La majorité des estimations indique entre 40 et 100 millions de locuteurs (dont moins de 5 millions sont des locuteurs natifs). Le swahili fait parti de la grande famille des langues bantu qui recouvre une large étendue géographique de plus de deux tiers du territoire d'Afrique Sub-Saharienne. En ce qui concerne la structure de la langue, le swahili est considéré comme une langue agglutinante (Marten, 2006). Elle possède les caractéristiques typiques des langues bantu comme les classes nominales et leurs systèmes d'accord et une morphologie verbale complexe. Cependant, elle se distingue de la plupart des autres langues bantu par l'absence de tons ainsi que par une part importante de son vocabulaire d'origine arabe. L'impact important du swahili en Afrique de l'Est explique pourquoi de nombreux acteurs principaux des services numériques proposent déjà une localisation de cette langue (entre autres : Wikipédia (2003), Google (2004), Microsoft (2005) et Facebook (2009)). De nombreuses autres initiatives pour la promotion du swahili sur le web existent. Sont à noter : le `Kamusi project` ("the internet living Swahili dictionary") ou le portail `goswahili.org` qui regroupe de considérables ressources sur la langue. Il faut aussi retenir le projet Kiswahili Linux Localization (`klnX`) qui a consacré des efforts importants à localiser des logiciels libres et ouverts en swahili.

En ce qui concerne le traitement automatique du langage naturel, des travaux antérieurs ont porté sur différents analyseurs (analyseur morphologique, segmenteurs, marqueurs de position, lemmatiseurs...). Certains utilisent une approche à base de règles comme dans (Hurskainen, 2004b), alors que d'autres favorisent une approche guidée par les données (De Pauw *et al.*, 2006; De Pauw et De Schryver, 2009; Shah *et al.*, 2010). Il est important de mentionner aussi les travaux en technologies du langage suivants : ceux en synthèse vocale (Ngugi *et al.*, 2010), en traduction automatique dans (De Pauw *et al.*, 2011a) et (De Pauw *et al.*, 2011b). Enfin, un premier système de dictée vocale est présenté en (Miriti, 2010).

Dans cette contribution, nous rapportons nos récents travaux sur le développement d'un système de reconnaissance automatique de la parole (RAP) pour le swahili. Dans la section suivante, nous présentons un aperçu de la situation linguistique et numérique de la langue. La section 3 retrace la collecte de données alors que la section 4 expose les résultats du système de RAP. Enfin, la section 5 conclue et discute des travaux à venir.

2 Ressources

2.1 Collecte et conception d'un corpus de texte

Un corpus de texte est indispensable pour la modélisation du langage en RAP. Des études récentes ont porté sur la collecte de textes en swahili : le corpus d'Helsinki (Hurskainen, 2004a) contient 12M de mots, (De Pauw *et al.*, 2011a) développent un corpus parallèle anglais-swahili de 2M de mots et enfin, 5M de mots sont collectés dans (Getao et Miriti, 2006). Le swahili bénéficiant d'une bonne visibilité sur le web, il a été décidé de construire notre propre corpus basé sur 16 sites d'information présélectionnés pour être strictement monolingue (évitant ainsi une étape de filtrage multilingue). De manière similaire à (Le *et al.*, 2003), toutes les pages d'articles

d'information ont été téléchargées sous différents formats, auxquelles ont été appliqués les processus d'extraction de texte, nettoyage et filtrage. À travers ce processus, plus de 28M de mots (tokens) ont été collectés.

Comme il a été décrit en section 1, le swahili est une langue agglutinante possédant une morphologie riche. Dans la structure verbale d'un verbe swahili, dix positions peuvent être identifiées (Marten, 2006). Si toutes ne peuvent être remplies en même temps, il est fréquent de trouver six ou sept positions remplies comme dans l'exemple suivant : `hawatakuambieni` est segmenté `ha-wa-ta-ku-ambi-e-ni` et glosé `NEG-SM2-FUT-OM2-tell-FIN-PL` [1]. De telles caractéristiques morphologiques impliquent une importante variété léxicale. Pour la RAP, cela entraine un manque de données et une couverture lexicale bien plus mauvaise que l'état de l'art actuel des système de RAP (comme on peut trouver pour l'anglais). Le considérable taux de mot hors vocabulaire (HV) qui en découle a des conséquences évidentes sur le taux d'erreur de mots (%Err) d'un système. Effectivement, chaque mot HV ne sera pas reconnu mais influera aussi la reconnaissance des mots voisins avec, comme impact immédiat, une montée du taux d'erreur. De nombreuses recherches se sont portées sur le traitement des langues à morphologie riche en TALN (Sarikaya *et al.*, 2009). En RAP, une solution est d'atteindre une couverture lexicale plus large en segmentant les mots en sous-unités, comme dans (Pellegrini et Lamel, 2009) pour l'amharique. Il est présenté dans (Hirsimaki *et al.*, 2009) un récent tour d'horizon de différentes études sur les modèles de langage basé sur le morphe [2] en RAP. Après avoir étudié différentes sous-unités pour le swahili (expérimentations non-reportées ici par manque de place), le morphe obtenu par une approche non-supervisée a été retenu. Pour ceci, nous avons utilisé l'outil publiquement disponible Morfessor (Creutz et Lagus, 2005). Il s'agit d'une approche guidée par les données qui apprend un lexique de sous-mots en utilisant un algorithme de minimisation de la taille de description (Minimum Description Length) à partir d'un corpus d'entrainement de mots. Si l'on considère le pourcentage de types HV selon le niveau de segmentation et différentes tailles de vocabulaire, la segmentation en morphes permet d'atteindre une couverture lexicale bien meilleure tout en gardant la même taille de vocabulaire : 19.17% de types HV avec un lexique de 65k mots et 11.36% avec 65k morphes. Dû aux limites du décodeur, cette étude se restreint à un vocabulaire de 65k. Néanmoins, pour un lexique de 200k mots, le taux de types HV est de 12.46% et toujours de 10.28% avec 400k mots (l'ensemble des mots disponibles). En parallèle, croître la taille du lexique à 200k morphes serait bien plus avantageux car il permettrait un taux de types HV de 1.61%.

2.2 Dictionnaire de prononciation

Le dictionnaire de prononciation est un élément primordial de la modélisation acoustique. Afin de le générer, nous avons extrait du corpus de texte les 65k mots les plus fréquents. L'étape suivante est de fournir une prononciation pour chacune des entrées lexicales en utilisant un nombre limité de phones, l'unité de base des modèles acoustiques. L'orthographe swahili est très proche de sa prononciation et très régulier : pour chaque phonème, l'unité de base linguistique, une seule même forme écrite est adoptée. Par conséquent, un script graphème vers phonème tire pleinement bénéfice de cette régularité et permet de générer la majeure partie des prononciations.

1. NEG= Negation, SM2= Marque sujet de la classe nominale 2 (il s'agit d'une des 16 différentes classes, il est fréquent en linguistique bantu de nommer ces classes nominales selon un système numérique), FUT= Temps futur, OM2= Marque objet de la classe nominale 2, *tell*= Racine verbale, FIN= Voyelle finale, PL= Pluriel post-finale

2. Le terme morphe est utilisé ici pour cette unité entre la syllabe et le mot. Selon le type de segmentation, elle peut correspondre au morphème, unité minimale porteuse de sens. Mais dans certain cas, avec une segmentation non-supervisée, elle peut ne correspondre à aucun type d'unité linguistique.

L'ensemble des phonèmes de la langue sont ici considérés comme phones, cependant, une analyse plus approfondie est nécessaire afin de décider si les sons les plus rares pourraient être évités et ainsi améliorer ou non le modèle acoustique. Notre système de RAP pour le swahili comptabilise 37 phones.

La génération de prononciation pour les mots anglais est un problème qui subsiste, ainsi que pour les noms propres et acronymes qui apparaissent tous fréquemment dans le corpus. La grande majorité des mots anglais et noms propres sont prononcés dans les émissions d'information tels qu'ils le sont en anglais. Si ces mots demeurent trop rares pour ajouter des phones spécifiques à l'anglais dans le modèle acoustique, ils sont aussi trop fréquents pour les laisser ainsi avec une prononciation erronée due à la règle graphème-phonème (de la même manière que (Chang *et al.*, 2011) avec le mandarin). Dans notre lexique de 65k mots swahili, 8,77% des mots se retrouvent dans le dictionnaire anglais de prononciation publiquement disponible du CMU (Carnegie Mellon University). Ces mots sont en grande majorité des mots anglais ou des noms propres. Ensuite, lorsqu'un mot est commun à la fois au dictionnaire CMU et à notre vocabulaire de 65k mots, la prononciation CMU est rajoutée en tant que variante de prononciation à notre dictionnaire. Les phones anglais sont transposés à ceux du swahili en procédant au préalable à un mapping théorique. Par exemple, le terme 'ukraine' est initialement phonétisé sous la forme '`u k r a i n e`' et nous rajoutons à partir du dictionnaire CMU, la variante : '`y u k r e y n`'. En ce qui concerne la prononciation des acronymes, ils sont le plus fréquemment prononcés de manière épellée. Ainsi, afin de générer des transcriptions plus proches qu'avec la règle graphème-phonème, un script détecte les entrées courtes contenant des clusters de lettres non-admis dans la phonotactique du swahili. Pour ces entrées, une variante avec la prononciation épellée est ajoutée (ex. TFF devient dans notre dictionnaire "`t i e f e f`").

2.3 Corpus audio

Afin d'effectuer l'apprentissage des modèles acoustiques, il est nécessaire d'avoir des données audio ainsi que les transcriptions correspondantes. Cependant, dans une situation de langues peu dotées, il est commun de ne pas avoir accès à ces ressources, ce qui représente donc une contrainte majeure au déploiement d'un système de RAP (Barnard *et al.*, 2009). Il s'agit d'une tâche à la fois longue, répétitive et coûteuse. De nombreuses études ont proposé des méthodologies dans le but d'accélérer la création de tels corpus comme dans (Davel *et al.*, 2011) et (Hughes *et al.*, 2010). Pour le swahili, nous avons d'abord commencé par collecter un corpus de parole lue. Les enregistrements ont été faits par 5 locuteurs natifs (2 femmes et 3 hommes), totalisant ainsi 3 heures et demie de parole lue transcrites. Afin d'obtenir un corpus plus conséquent, nous avons aussi collecté plus de 200h d'émissions d'information radiodiffusées sur le web.

Dans le but de fournir rapidement les transcriptions de ce corpus, nous avons exploré l'usabilité de l'outil de crowdsourcing Amazon's Mechanical Turk (MTurk). Mturk est un marché de travail en ligne où quiconque peut soumettre de simples tâches à des personnes volontaires. De nombreuses études récentes ont démontré la pertinence et la puissance de cet outil pour des tâches de TALN (Parent et Eskenazi, 2011). Spécifiquement pour les transcriptions, il possède un grand potentiel à réduire le coût et le temps tout en gardant une qualité suffisante (Novotney et Callison-Burch, 2010). Mais une certaine polémique entre chercheurs entoure Mturk pour certaines raisons légales et éthiques (mentionnées dans (Gelas *et al.*, 2011) et (Adda *et al.*, 2011)). Pour cette raison, nous avons d'abord seulement évalué sur le petit corpus de parole lue la possibilité de l'utiliser. Le modèle acoustique appris avec les transcriptions MTurk était très proche de celui utilisant les transcriptions de référence. Respectivement 38.5% et 38% de %Err sont obtenues

sur un petit corpus de test de 82 phrases (détails dans (Gelas *et al.*, 2011) où le processus est aussi appliqué à l'amharique). La transcription de 3 heures et demie de parole lue s'est complétée en 12 jours par trois personnes sur MTurk. Il s'agit clairement d'un taux d'accomplissement plus faible que pour l'anglais. Ceci ajouté aux potentielles questions d'éthiques, nous avons décidé de travailler directement avec un institut kenyan [3] pour transcrire collaborativement 12 heures de notre corpus d'émissions d'information radiodiffusées.

L'optique principale est encore de faciliter et de réduire le temps pris par la transcription. Ainsi, nous avons considéré un processus de transcription collaboratif basé sur l'application itérative du protocole suivant : un premier modèle acoustique est appris en utilisant les données du corpus de parole lue. Ensuite, chaque émission est segmentée en utilisant une détection de silence automatique standard (seuls les fichiers dont la durée est entre 2 et 6 secondes sont gardés afin de pré-filtrer une partie des segments musicaux et trop bruités). Ensuite, un ensemble de deux heures d'audio présegmentées et pré-filtrées est transcrit par notre premier système de RAP. La sortie de ce décodage est envoyée à l'Institut Taji pour correction (post-edition). Enfin, après être corrigées par les transcripteurs, les données annotées sont ajoutées au corpus d'apprentissage et un nouveau modèle acoustique est entraîné dans le but de transcrire l'ensemble de deux heures suivant. Cette procédure est répétée jusqu'à ce que 12 heures de paroles transcrites soient obtenues, en gardant 10 heures pour l'apprentissage et 2 heures comme corpus de test. Il apparaît que le temps passé à post-éditer les transcriptions est corrélé avec la qualité des transcriptions pourvues. Les résultats du tableau 1 (du 1er au 6^{ième} set) montrent que chaque ensemble correctement transcrit rajouté au corpus d'apprentissage améliore le modèle acoustique. Celui-ci fournit donc de meilleures transcriptions pour la tranche audio suivante et demande par conséquent moins de temps à corriger. À l'aide de ce protocole, la durée de la tâche de transcription est passé de 40 heures (1er set) à environ 26 heures (2^{ième} au 5^{ième} set) pour enfin atteindre 15 heures (6^{ième} set).

3 Système de reconnaissance automatique de la parole

3.1 Configuration du système

Une fois toutes les ressources décrites auparavant collectées, nous avons utilisé la boîte à outils SphinxTrain [4] afin de développer les modèles acoustiques (MA) à base de modèles de Markov cachés à 3 états pour le swahili. L'étape initiale est d'extraire les paramètres acoustiques via une fenêtre glissante. Chaque trame a une taille de 25ms dont le début est incrémenté de 10ms. Le signal audio est ainsi paramétré selon 13 coefficients MFCC (Mel Frequency Cepstral Coefficients). Ensuite, ces paramètres acoustiques permettent l'apprentissage d'un modèle dépendant du contexte (CD) (3000 états). Pendant le travail de transcription collaboratif, seuls des modèles indépendants du contexte (CI) sont appris jusqu'à que 10 heures de données audio d'apprentissage soient atteintes. En ce qui concerne les modèles de langage, autant les trigrammes à base de mots que de morphes sont construits à l'aide de la boîte à outils du SRI [5].

3.2 Résultats

Différentes expérimentations de RAP ont été conduites sur un corpus de test de 2 heures (1991 phrases) et les résultats sont présentés tableau 1. Comme attendu lors du travail de

3. http ://www.taji-institute.com/
4. cmusphinx.sourceforge.net/
5. www.speech.sri.com/projects/srilm/

transcription collaboratif, chaque ensemble de 2 heures ajouté aux données d'apprentissage améliore significativement (exception faite entre le 4^{ième} et le 5^{ième}) le taux d'erreur. Le passage visible d'un modèle acoustique CI vers CD était notre dernière étape pour notre modèle de référence, atteignant ainsi 35.8 %Err.

Dans ce même tableau, il est aussi possible de voir comment l'ajout des variantes de prononciation pour l'anglais et les acronymes augmente les performances dans un environnement acoustique propre (26.9 %Err sans et 26.5 %Err avec). Lorsque l'on considère les qualités audio plus dégradées, elles amènent une trop grande confusion dans le signal pour être véritablement bénéfiques (de 35.8 %Err à 35.7 %Err).

Finalement, dans notre expérience sur les sous-unités du mot au niveau du modèle de langage, les sorties du décodeur de RAP sont une séquence de morphes impliquant la nécessité de reconstruire le niveau mot. En conséquent, une balise de frontière de morphe est ajoutée de chaque côté de la segmentation. Pour reconstruire les sorties au niveau mot, nous reconnectons les unités chaque fois que deux frontières de morphes apparaissent consécutivement (exemple, `kiMB MBtabu` devient `kitabu`). L'utilisation de sous-unités au mot pour la modélisation du langage réduit significativement le taux d'erreur aussi bien dans un environnement acoustique bon que mauvais (34.8 %Err toutes qualités audio confondues et 25.9 %Err avec seulement la qualité studio). Ceci peut être expliqué par l'augmentation de la couverture lexicale. La couverture du vocabulaire de 65k morphes représente 30.83% de l'ensemble du lexique quand le vocabulaire de 65k mots représente lui seulement 13.95%. Comme présenté en 2.1, ceci a un impact direct sur les mots HV. Effectivement, une autre qualité d'un modèle de langage basé sur des sous-unités pour la RAP est la récupération des mots initialement HV. Parmi les mots HV qui peuvent être reconnus, 36,04% sont retrouvés.

TABLE 1 – Taux d'erreur (%Err) selon les différents modèles acoustique (CI ou CD), modèles de langage (Mots ou Morphes), dictionnaire de prononciation (avec ou sans variantes) et la qualité audio (tout, téléphonique, bruité ou studio)

Système de RAP	Qualité audio	Nombre de phrases	%Err
1er Set CI Mot(65k)	Tout	1991	72.8
2^{ième} Set	Tout	1991	59.0
3^{ième} Set	Tout	1991	57.4
4^{ième} Set	Tout	1991	56.2
5^{ième} Set	Tout	1991	56.1
Référence CD Mot(65k)	Tout	1991	35.8
	Téléphonique	424	60.0
	Bruitée	402	36.4
	Studio	1165	26.9
Référence + Variantes dict	Tout	1991	35.7
	Studio	1165	26.5
CD Morphe(65k + variantes)	Tout	1165	**34.8**
	Studio	1165	**25.9**

4 Conclusion

Dans la présente contribution, il est décrit un ensemble de nouvelles ressources développées pour la RAP du swahili. Différentes approches pour accélérer la création d'un corpus de parole transcrit ont été explorées. MTurk, l'outil puissant de crowdsourcing, a été envisagé dans le but de pourvoir les transcriptions de notre corpus principal. Et même si sur un petit corpus controllé, l'expérience s'est avérée concluante, un processus de transcription collaboratif avec un institut kenyan a été préféré. Afin d'aider les transcripteurs dans leur tâche, une pré-transcription d'un ensemble de deux heures de parole leur était soumis à corriger. Une fois finalement correctement transcrites, les données étaient rajoutées au corpus d'apprentissage et un nouveau modèle acoustique était réappris améliorant ainsi les transcriptions proposées suivantes. Ce protocole a permis de réduire la durée d'annotation pour deux heures de parole de 40h à 15h.

Une attention particulière a aussi été portée sur certaines singularités linguistiques du swahili, en gardant à l'esprit la possibilité de reproduire ces méthodologies sur d'autres langues linguistiquement similaires. En ce qui concerne la modélisation du langage d'une langue à morphologie riche, l'utilisation de sous-unités au mot a permis d'améliorer les performances de notre système en appliquant des méthodes non-suppervisées. À l'aide de la segmentation proposée par Morfessor, nous sommes passés de 35.7 %Err pour le modèle de mot à 34.8 %Err avec le modèle de morphes. Une expérimentation parallèle portée sur une tâche de parole lue sur l'amharique s'est aussi montrée profitable, les résultats étant présentés dans (Tachbelie *et al.*, 2012).

Pour ce qui est du développement du dictionnaire de prononciation, la présence importante de termes anglais a été prise en compte. Des variantes de prononciation on été automatiquement générées en tirant avantage de matériaux déjà disponible comme le dictionnaire de prononciation anglais de CMU. Ce procédé associé à la détection automatique et génération de variantes pour les acronymes a permis d'améliorer les performances, notamment dans un environnement audio clair (qualité studio) en passant de 26.9 %Err à 26.5 %Err.

Références

Adda, G., Sagot, B., Fort, K. et Mariani, J. (2011). Crowdsourcing for language resource development : Critical analysis of amazon mechanical turk overpowering use. *In LTC, 5th Language and Technology Conference*.

Barnard, E., Davel, M. et Heerden, C. (2009). Asr corpus design for resource-scarce languages. *In Interspeech*.

Barnard, E., Schalkwyk, J., van Heerden, C. et Moreno, P. (2010). Voice search for development. *In Interspeech*.

Chang, H., Sung, Y., Strope, B. et Beaufays, F. (2011). Recognizing english queries in mandarin voice search. *In ICASSP*. IEEE.

Creutz, M. et Lagus, K. (2005). Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Rapport technique, Computer and Information Science, Report A81, Helsinki University of Technology.

Davel, M., van Heerden, C., Kleynhans, N. et Barnard, E. (2011). Efficient harvesting of internet audio for resource-scarce asr. *In Interspeech*.

De Pauw, G. et De Schryver, G. (2009). African language technology : The data-driven perspective. *V. Lyding (eds.)*, pages 79–96.

De Pauw, G., De Schryver, G. et Wagacha, P. (2006). Data-driven part-of-speech tagging of kiswahili. *In Text, speech and dialogue*, pages 197–204. Springer.

De Pauw, G., Wagacha, P. et De Schryver, G. (2011a). Exploring the sawa corpus : collection and deployment of a parallel corpus english - swahili. *Language resources and evaluation*, pages 1–14.

De Pauw, G., Wagacha, P. et de Schryver, G. (2011b). Towards english-swahili machine translation. *In Research Workshop of the Israel Science Foundation*.

Gelas, H., Abate, S., Besacier, L. et Pellegrino, F. (2011). Evaluation of crowdsourcing transcriptions for african languages. *In HLTD*.

Getao, K. et Miriti, E. (2006). Automatic construction of a kiswahili corpus from the world wide web. *Measuring Computing Research Excellence and Vitality*, page 209.

Hirsimaki, T., Pylkkonen, J. et Kurimo, M. (2009). Importance of high-order n-gram models in morph-based speech recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(4):724–732.

Hughes, T., Nakajima, K., Ha, L., Vasu, A., Moreno, P. et LeBeau, M. (2010). Building transcribed speech corpora quickly and cheaply for many languages. *In INTERSPEECH*.

Hurskainen, A. (2004a). Hcs 2004–helsinki corpus of swahili. *Compilers : Institute for Asian and African Studies (University of Helsinki) and CSC*.

Hurskainen, A. (2004b). Swahili language manager : a storehouse for developing multiple computational applications. *Nordic Journal of African Studies*, 13(3):363–397.

Kumar, A., Tewari, A., Horrigan, S., Kam, M., Metze, F. et Canny, J. (2011). Rethinking speech recognition on mobile devices. *In IUI4DR*. ACM.

Le, V., Bigi, B., Besacier, L. et Castelli, E. (2003). Using the web for fast language model construction in minority languages. *In Eighth European Conference on Speech Communication and Technology*.

Marten, L. (2006). Swahili. *In* Brown, K., éditeur : *The Encyclopedia of Languages and Linguistics, 2nd ed.*, volume 12, pages 304–308. Oxford : Elsevier.

Miriti, E. (2010). *A Kiswahili Dictation System : Implementation of a Prototype*. VDM Verlag Dr. Müller.

Ngugi, K., Okelo-Odongo, W. et Wagacha, P. (2010). Swahili text-to-speech system. *African Journal of Science and Technology*, 6(1).

Novotney, S. et Callison-Burch, C. (2010). Cheap, fast and good enough : Automatic speech recognition with non-expert transcription. *In NAACL HLT*, pages 207–215. Association for Computational Linguistics.

Parent, G. et Eskenazi, M. (2011). Speaking to the crowd : looking at past achievements in using crowdsourcing for speech and predicting future challenges. *In Interspeech*.

Patel, N., Chittamuru, D., Jain, A., Dave, P. et Parikh, T. (2010). Avaaj otalo : a field study of an interactive voice forum for small farmers in rural India. *In CHI*, pages 733–742. ACM.

Pellegrini, T. et Lamel, L. (2009). Automatic word decompounding for ASR in a morphologically rich language : Application to Amharic. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):863–873.

Polomé, E. (1967). *Swahili Language Handbook*. Center for Applied Linguistics, Washington, DC.

Sarikaya, R., Kirchhoff, K., Schultz, T. et Hakkani-Tur, D. (2009). Introduction to the special issue on processing morphologically rich languages. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5).

Shah, R., Lin, B., Gershman, A. et Frederking, R. (2010). Synergy : a named entity recognition system for resource-scarce languages such as swahili using online machine translation. *In Proceedings of the Second Workshop on African Language Technology (AfLaT 2010)*, pages 21–26.

Tachbelie, M. Y., Abate, S. T., Besacier, L. et Rossato, S. (2012). Syllable-based and hybrid acoustic models for amharic speech recognition. *In SLTU*.