Analyse des performances de modèles de langage sub-lexicale pour des langues peu-dotées à morphologie riche

Hadrien Gelas[1,2] Solomon Teferra Abate[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
{solomon.abate, laurent.besacier@imag.fr}

RÉSUMÉ

Ce papier étudie l'impact de l'utilisation d'unités sous-lexicales sur les performances d'un système de RAP pour deux langues africaines peu-dotées et morphologiquement riches (l'amharique et le swahili). Deux types de sous-unités sous-lexicales sont considerés : la syllabe et le morphème, ce dernier étant obtenu de manière supervisée ou non-supervisée. La reconstruction en mots à partir de sorties de RAP en syllabes ou morphèmes est aussi prise en compte. Pour les deux langues, les meilleurs résultats sont obtenus avec les morphèmes non-supervisés. Le taux d'erreur de mots est grandement réduit pour la reconnaissance de l'amharique dont les données d'entrainement du LM sont très faibles (2,3M de mots). Les scores pour la RAP du swahili sont aussi améliorés (28M de mots pour l'entrainement). Il est ausi présenté une analyse détaillée de la reconstruction des mots hors vocabulaires, un pourcentage important de ceux-ci (jusqu'à 75% pour l'amharique) sont retrouvés à l'aide de modèles de langage à base de morphèmes et la méthode de reconstruction appropriée.

ABSTRACT

Performance analysis of sub-word language modeling for under-resourced languages with rich morphology : case study on Swahili and Amharic

This paper investigates the impact on ASR performance of sub-word units for two under-resourced african languages with rich morphology (Amharic and Swahili). Two subword units are considered : syllable and morpheme, the latter being obtained in a supervised or unsupervised way. The important issue of word reconstruction from the syllable (or morpheme) ASR output is also discussed. For both languages, best results are reached with morphemes got from unsupervised approach. It leads to very significant WER reduction for Amharic ASR for which LM training data is very small (2.3M words) and it also slightly reduces WER over a Word-LM baseline for Swahili ASR (28M words for LM training). A detailed analysis of the OOV word reconstruction is also presented ; it is shown that a high percentage (up to 75% for Amharic) of OOV words can be recovered with morph-based language model and appropriate reconstruction method.

MOTS-CLÉS : Modèle de langage, Morphème, Hors vocabulaire, Langues peu-dotées.

KEYWORDS: Language model, Morpheme, Out-of-Vocabulary , Under-resourced languages.

JEP-TALN-RECITAL 2012, Atelier TALAf 2012: Traitement Automatique des Langues Africaines, pages 53–62,
Grenoble, 4 au 8 juin 2012.

1 Introduction

Due to world's globalisation and answering the necessity of bridging the numerical gap with the developing world, speech technology for under-resourced languages is a challenging issue. Applications and usability of such tools in developing countries are proved to be numerous and are highlighted for information access in Sub-Saharan Africa (Barnard *et al.*, 2010a,b), agricultural information in rural India (Patel *et al.*, 2010), or health information access by community health workers in Pakistan (Kumar *et al.*, 2011).

In order to provide a totally unsupervised and language independent methodology to develop an automatic speech recognition (ASR) system, some particular language characteristics should be taken into account. Such specific features as tones ((Lei *et al.*, 2006) on Mandarin Chinese) or writing systems without explicit word boundaries ((Seng *et al.*, 2008) on Khmer) need a specific methodology adaptation. This is especially true when dealing with under-resourced languages, where only few data are available.

During recent years, many studies tried to deal with morphologically rich languages (whether they are agglutinative, inflecting and compounding languages) in NLP (Sarikaya *et al.*, 2009). Such a morphology results in data sparsity and in a degraded lexical coverage with a similar lexicon size than state-of-the-art speech recognition setup (as one for English). It yields high Out-of-Vocabulary (OOV) rates and degrades Word-Error rate (WER) as each OOV words will not be recognized but can also affect their surrounding words and strongly increase WER.

When the corpus size is limited, a common approach to overcome the limited lexical coverage is to segment words in sub-word units (morphemes or syllables). Segmentation in morphemes can be obtained in a supervised or unsupervised manner. Supervised approaches were mainly used through morphological analysers built on carefully annotated corpora requiring important language-specific knowledge (as in (Arısoy *et al.*, 2009)). Unsupervised approaches are language-independent and do not require any linguistic-knowledge. In (Kurimo *et al.*, 2006), several unsupervised algorithms have been compared, including their own public method called Morfessor ((Creutz et Lagus, 2005)) for two ASR tasks in Turkish and Finnish (see also (Hirsimaki *et al.*, 2009) for a recent review of morh-based approaches). The other sub-word type that is also utilized for reducing high OOV rate is the syllable. Segmentation is mainly rule-based and was used in (Shaik *et al.*, 2011b) and (Shaik *et al.*, 2011a), even if outperformed in WER by ASR morpheme-based recognition for Polish and German.

In this work, we investigate those different methodologies and see how to apply them for two different speech recognition tasks : read speech ASR in Amharic and broadcast speech transcription in Swahili. These tasks represents two different profiles of under-resourced languages cases. Amharic with an acoustic model (AM) trained on 20h of read-speech but limited text data (2.3M) and on the opposite, Swahili with a weaker acoustic model (12h of broadcast news from internet mixing genre and quality) but a more robust LM (28M words of web-mining news, still without any adaptation to spoken broadcast news). If such study on sub-unit has already been conducted on Amharic (Pellegrini et Lamel, 2009), no prior work are known to us for Swahili. But, the main goal of this study is to better understand what does really impact performance of ASR using sub-word unit through a comparison of different methodologies. Both supervised and unsupervised segmentation strategies are explored as well as different approaches to tag segmentation.

The next section describes the target languages and the available corpora. Then, we introduce several segmentation approaches in section 3. Section 4 presents the analysis of experimental results for Swahili and Amharic while section 5 concludes this work.

2 Experiment description

2.1 Languages

Amharic is a Ethio-Semitic language from the Semitic branch of the Afroasiatic super family. It is related to Hebrew, Arabic, and Syrian. According to the 1998 census, it is spoken by over 17 million people as a first language and by over 5 million as a second language throughout Ethiopia. Amharic is also spoken in other countries such as Egypt, Israel and the United States. It has its own writing system which is syllabary. It exhibits non-concatenative, inflectional and derivational morphology. Like other Semitic languages such as Arabic, Amharic exhibits a root-pattern morphological phenomenon. Case, number, definiteness, and gender-marking affixes inflect nouns. Some adverbs can be derived from adjectives but adverbs are not inflected. Nouns are derived from other basic nouns, adjectives, stems, roots, and the infinitive form of a verb is obtained by affixation and intercalation.

Swahili is a Bantu language often used as a vehicular language in a wide area of East Africa. It is not only the national language of Kenya and Tanzania but it is also spoken in different parts of Democratic Republic of Congo, Mozambique, Somalia, Uganda, Rwanda and Burundi. Most estimations give over 50 million speakers (with only less than 5 million native speakers). It has many typical Bantu features, such as noun class and agreement systems and complex verbal morphology. Structurally, it is often considered as an agglutinative language (Marten, 2006).

2.2 Speech corpora description

Both Amharic and a small part of Swahili training audio corpora were collected following the same protocol. Texts were extracted from news websites and segmented by sentence. Native speakers were recorded using a self-paced reading interface (with possible rerecordings). The Amharic speech corpus (Abate *et al.*, 2005) consists of 20 hours of training speech collected from 100 speakers who read a total of 10,850 sentences. Swahili corpus corresponds to 2 hours and a half read by 5 speakers (3 male and 2 female) along with almost 10 hours of web-mining broadcast news representing various types of recording quality (noisy speech, telephone speech, studio speech) and speakers. They were transcribed using a collaborative transcription process based on the use of automatic pre-transcriptions to increase productivity gains (See details in (Gelas *et al.*, 2012)). Test corpora are made of 1.5 hours (758 sentences) of read speech for Amharic and 2 hours (1,997 sentences) of broadcast news for Swahili.

2.3 Text corpora description

We built all statistical N-gram language model (LM) using the SRI [1] language model toolkit. Swahili text corpus is made of data collected from 12 news websites (over 28M words). To generate a pronunciation dictionary, we extracted the 65k most frequent words from the text corpus and automatically created pronunciations taking benefit of the regularity of the grapheme to phoneme conversion in Swahili. The same methodology and options have been applied to all sub-words LM. For Amharic, we have used the data (2.3M words text) described in (Tachbelie *et al.*, 2010).

3 Segmenting text data

3.1 Unsupervised morphemic segmentation

For the unsupervised word segmentation, we used a publicly available tool called Morfessor [2]. Its data-driven approach learns a sub-word lexicon from a training corpus of words by using a Minimum Description Length (MDL) algorithm (Creutz et Lagus, 2005). It has been used with default options and without any adaptation.

3.2 Supervised morphemic and syllabic segmentation

For Amharic, we used the manually-segmented text described in (Tachbelie *et al.*, 2011a) to train an FSM-based segmenter (a composition of morpheme transducer and 12gram consonant vowel syllable-based language model) using the AT&T FSM Library (FiniteState Machine Library) and GRM Library (Grammar Library)(Mohri *et al.*, 1998). The trained segmenter with the language model is applied to segment the whole text mentioned in (Tachbelie *et al.*, 2010).

The supervised decomposition for Swahili is performed with the public Part-Of-Speech tagger named TreeTagger [3]. It is using the parameters available for Swahili to extract sub-word units.

As for as syllable segmentation is concerned, we designed rule-based algorithms following structural and phonological restrictions of the respective languages.

3.3 Segmentation tagging and vocabulary size

While working on sub-word unit, one should think on how to incorporate the segmentation information. Morphological information can be included within factored LM as in (Tachbelie *et al.*, 2011b) or directly as a full unit in the n-gram LM itself. By choosing the latter, the ASR decoder output is a sequence of sub-word units and an additional step is needed to recover

1. www.speech.sri.com/projects/srilm/
2. The unit obtained with Morfessor is referred here as morpheme even if it do not automatically corresponds to the linguistic definition of morpheme (the smallest semantically meaningful unit)
3. www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTree Tagger.html

words from sub-units. In (Diehl *et al.*, 2011), a n-gram SMT-based morpheme-to-word conversion approach is proposed.

In this work, we evaluate how the recognition performance is affected by different ways of tagging the segmentation information straightly in the training text. In (Arısoy *et al.*, 2009), it is noticed that this aspect need to be considered as it impacts WER. In (Guijarrubia *et al.*, 2009), a similar methodology is applied without reading any conclusion since a too small and easy recognition task was performed.

Three distinct types of tagging are evaluated here :

– UNIT_AFX : A morpheme boundary (MB) is added on left and/or right side of segmentation leaving the (so-called with Morfessor) root alone. To rebuild up to words, we reconnect every units containing MB to the one next to it.
(ex. `kiMB tabu` → `kitabu`)
– UNIT_ALL : A MB tag is added on each side of segmentation, in other words, we add to the lexicon the information to distinguish roots from their context (we can get up to four different entries for a same root : ROOT, MBROOT, ROOTMB, MBROOTMB). To rebuild, we reconnect every time two MB appearing consecutively.
(ex. `kiMB MBtabu` → `kitabu`)
– UNIT_POS : For syllables, we add to the unit the position of the syllable in the word.
(ex. `1ki 2ta 3bu` → `kitabu`)

In table 1, it is shown that each choice has an influence on the size of the full text vocabulary and thus on the lexical coverage of the 65k lexicon. As expected from a language with rich morphology, the word baseline 65k lexicon shows a dramatically low lexical coverage (13.95%). For the same text information, syllables logically reduce the size of vocabulary and got a full theoretical lexical coverage without reaching the 65k limits, but with the cost of really short length unit. Concerning both morpheme segmentation types, as expected the supervised approach leads to a larger number of units than the unsupervised statistical approach, the latter leads to a better theoretical lexical coverage. The average token length do not reduce much compared to word unit as most frequent words are already short mono-morphemic grammatical words. The influence of different tagging techniques is also shown on the same table. Detailed comments on WER will be given in 4.2.

LM	FullVoc (%)	65k Cov. (%)	Token length	WER (%)
Word	100	13.95	5.5	35.7
Syl_Pos (V=27k)	5.79	100	2.0	51.7
Treetag_All	79.38	17.57	4.4	44.7
Treetag_Afx	78.61	17.74	4.4	43.3
Morf_All	45.24	30.83	5.3	**34.8**
Morf_Afx	38.07	36.64	5.3	35.4

TABLE 1 – *Swahili - Size of full text corpus vocabulary in comparison with a word level baseline (FullVoc) ; lexical coverage of a 65k lexicon on the full vocabulary (65k Cov.) ; average token length in character for the text corpus ; word error rate depending on the choice of unit and segmentation tag (WER), all systems using 3gram LM and 65k lexicon except when specified*

4 Results

4.1 ASR system description

We used SphinxTrain [4] toolkit from Sphinx project for building Hidden Markov Models (HMM) based acoustic models (AMs) for Swahili. With the speech training database described in 2.2, we trained a context-dependent model with 3,000 tied states. The acoustic models have 36 and 40 phones for Swahili and Amharic, respectively. We used the HDecode decoder of the HTK for Amharic. The Amharic acoustic model is more precisely described in (Tachbelie *et al.*, 2010).

4.2 Analysis of Sub-word units performance for Swahili

Comparing all results for Swahili broadcast speech transcription task (table 1), Morfessor based segmentation ASR system is the only one, with 34.8% WER, performing significantly better than the 35.7% word baseline. As in (Arısoy *et al.*, 2009) and (Hirsimaki *et al.*, 2006), segmentation based on a morphological analyser reaches lower results (43.3% WER) than words and unsupervised based segmentation. Finally, rule-based syllabic system have the worst performance with 51.7% WER. Those scores in table 1 gives a good indication on how to choose the most performing unit. It seems that one need to balance and optimise two distinct criteria : n-gram length coverage and lexical coverage.

The importance of n-gram length coverage can be seen with poor performance of too short units, like syllables in this work. A syllable trigram (average 6.0 character-long) is approximately equivalent to a word unigram in Swahili (average 5.5 character-long), thus such a short trigram length is directly impacting ASR system performance even if lexical coverage is maximized (100%). The importance to use higher order n-gram LM when dealing with short units is also shown in (Hirsimaki *et al.*, 2009). However, if a lattice rescoring framework is often used, it is difficult to recover enough information if the first trigram pass do not perform well enough. It is then recommended to directly implement the higher order n-gram LM in the decoder.

In the same time, a larger lexical coverage (lex.cov.), allows better performance if not used with too short units as shows the difference of performance between word-based LM (13.95% lex.cov. and 35.7% WER) and Morfessor-based LM (30.83% lex.cov. and 34.8% WER), both having similar average token lengths.

Concerning the different tagging techniques, they have an impact on WER. The better choice seems to be influenced by the lexical coverage. When lexical coverage is good enough (Morfessor-based system), one can get advantage of having more different and precise contexts (tag on all units, separating roots alone and roots with affixes in the vocabulary and on n-gram estimations), whereas for low lexical coverage (TreeTagger-based system), having more various words is better (tag only on affixes, regrouping all same roots together allowing more distinct units in the lexicon).

4. cmusphinx.sourceforge.net/

4.3 Sub-word units performance for Amharic

For the read speech recognition task for Amharic, only the best performing systems are presented in table 2. Similar trend is found concerning the tagging techniques (better systems are tagged ALL for Morfessor and tagged AFX for FSM) and by the fact that Morfessor system outperforms the others. Even if the unit length in Morfessor is 40% shorter than average word length, it gets important benefits from a 100% lexical coverage of the training corpus. However, for this task, the supervised segmentation (FSM) has better results than word baseline system. It can be explained by a slightly increased lexical coverage and still a reasonable token length. Through this task, we also considered several vocabulary sizes. Results show that WER greatly benefits from sub-units in smaller lexicon tasks. Finally, as for Amharic sub-word units being notably shorter than word units, we rescored output lattices from the trigram LM system with a 5gram LM. It leads to an absolute WER decrease of 2.0% for Morfessor.

	65k Cov.	Token	Word Error Rate (%)		
LM	(%)	length	5K	20K	65K
Word_3g	30.79	8.3	52.4	29.6	15.9
FSM_Afx_3g	45.13	6.3	39.3	20.8	12.2
FSM_Afx_5g	45.13	6.3	39.1	20.3	11.4
Morf_All-3g	100	4.9	36.7	14.8	9.9
Morf_All-5g	100	4.9	34.9	12.6	**7.9**

TABLE 2 – *Amharic - Lexical coverage of a 65k lexicon on the full vocabulary (65k Cov.) ; average token length in the whole text corpus ; word error rate depending on the choice of unit, segmentation tag and vocabulary size*

4.4 OOV benefits of using sub-word units

Making good use of sub-word units for ASR has been proved efficient in many research to recognize OOV words over baseline word LMs (as in (Shaik *et al.*, 2011a)). Table 3 presents the different OOV rates considering both token and type for each LM (OOV morphemes for Morfessor-based LM). We also present the proportion of correctly recognized words (COOV) which were OOVs in the word baseline LM. Results show important OOV rate reduction and correctly recognised OOV rate for both languages (Morfessor-based outputs). For Amharic, the difference of COOV rate between each lexicon is correlated with the possible OOVs each system can recognized.

Swahili obtain less benefits for COOV. It can be explained by the specificity of the broadcast news task, leading to important OOV entity names or proper names (the 65k Morfessor-based lexicon is still having 11.36% of OOV types). But if we consider only the OOVs that can possibly be recognized (i.e. only those which are not also OOVs in the Morfessor-based lexicon), 36.04% of them are rebuilt. Due to decoder limitations we restrained this study to a 65k lexicon, but for a Swahili 200k word vocabulary we get a type OOV rate of 12.46% and still 10.28% with a full vocab (400k). Those numbers are really close to those obtained with the 65k Morfessor lexicon and could only be reached with the cost of more computational power and less robust LM. In the

LM	OOV (%) Token	OOV (%) Type	COOV (%)
Amharic			
Word-5k	35.21	57.14	-
Word-20k	19.48	32.18	-
Word-65k	9.06	14.99	-
Morf_All-5k	13.67	40.58	33.76
Morf_All-20k	2.50	7.88	66.95
Morf_All-65k	0.12	2.81	**75.30**
Swahili			
Word-65k	5.73	19.17	-
Morf_All-65k	3.67	11.36	8.77

TABLE 3 – *Amharic and Swahili - Token and type OOV rate in test reference transcriptions depending on LM (OOV morphemes for Morfessor-based LM) ; correctly recognised baseline OOV words rate in ASR outputs (COOV)*

same time, growing Morfessor lexicon to 200k would be more advantageous as it reduces the type OOV rate to 1.61%.

While using sub-word system outputs rebuilt to word level reduces OOV words, in contrary, it can also generate non words by ungrammatical or non-sense concatenation. We checked the 5029 words generated by the best Amharic Morfessor output to see if they exist in the full training text vocabulary. It appears that only 37 are non-words (33 after manual validation). Among those 33, there were 26 isolated affixes and 7 illegal concatenations, all due to poor acoustic estimation from the system. Considering this small amount of non-words and with no possibility to retrieve good ones in lattices, we did not process to constraint illegal concatenation as in (Arısoy et Saraçlar, 2009).

5 Conclusion

We investigated the use of sub-word units in n-gram language modeling through different methodologies. The best results are obtained using unsupervised segmentation with Morfessor. This tool outperforms supervised methodologies (TreeTagger, FSM or rule-based syllables) because the choice of sub-word units optimise two essential criteria which are n-gram length coverage and lexical coverage. In the same time, it appears that the way one implements the segmentation information affects the speech recognition performance. As expected, using sub-word units brings major benefits to the OOV problem. It shows to be effective in two very different tasks for two under-resourced African languages with rich morphology (one being highly inflectional, Amharic and the other being agglutinative, Swahili). The Amharic read speech recognition task, get the more advantages of it, since the word baseline LM suffers from data sparsity. But results are also improved for a broadcast speech transcription task for Swahili.

Références

Abate, S., Menzel, W. et Tafila, B. (2005). An Amharic speech corpus for large vocabulary continuous speech recognition. *In Interspeech*, pages 67–76.

Arisoy, E., Can, D., Parlak, S., Sak, H. et Saraçlar, M. (2009). Turkish broadcast news transcription and retrieval. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(5):874–883.

Arisoy, E. et Saraçlar, M. (2009). Lattice extension and vocabulary adaptation for Turkish LVCSR. *Audio, Speech, and Language Processing, IEEE Transactions on*, 17(1):163–173.

Barnard, E., Davel, M. et van Huyssteen, G. (2010a). Speech technology for information access : a South African case study. *In AAAI Symposium on Artificial Intelligence*, pages 22–24.

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

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.

Diehl, F., Gales, M., Tomalin, M. et Woodland, P. (2011). Morphological decomposition in Arabic ASR systems. *Computer Speech & Language*.

Gelas, H., Besacier, L. et Pellegrino, F. (2012). Developments of swahili resources for an automatic speech recognition system. *In SLTU*.

Guijarrubia, V., Torres, M. et Justo, R. (2009). Morpheme-based automatic speech recognition of basque. *Pattern Recognition and Image Analysis*, pages 386–393.

Hirsimaki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S. et Pylkkonen, J. (2006). Unlimited vocabulary speech recognition with morph language models applied to Finnish. *Computer Speech & Language*.

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.

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

Kurimo, M., Creutz, M., Varjokallio, M., Arisoy, E. et Saraclar, M. (2006). Unsupervised segmentation of words into morphemes–morpho challenge 2005, application to automatic speech recognition. *In Interspeech*.

Lei, X., Siu, M., Hwang, M., Ostendorf, M. et Lee, T. (2006). Improved tone modeling for Mandarin broadcast news speech recognition. *In Interspeech*.

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

Mohri, M., Pereira, F. et Riley, M. (1998). A rational design for a weighted finite-state transducer library. *In Lecture Notes in Computer Science*, pages 144–158. Springer.

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.

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).

SENG, S., SAM, S., BESACIER, L., BIGI, B. et CASTELLI, E. (2008). First broadcast news transcription system for Khmer language. *In LREC*.

SHAIK, M., MOUSA, A., SCHLUTER, R. et NEY, H. (2011a). Hybrid language models using mixed types of sub-lexical units for open vocabulary German LVCSR. *In Interspeech*.

SHAIK, M., MOUSA, A., SCHLUTER, R. et NEY, H. (2011b). Using morpheme and syllable based sub-words for Polish LVCSR. *In ICASSP*.

TACHBELIE, M., ABATE, S. et BESACIER, L. (2011a). Part-of-speech tagging for under-resourced and morphologically rich languages - the case of Amharic. *In HLTD*.

TACHBELIE, M., ABATE, S. et MENZEL, W. (2010). Morpheme-based automatic speech recognition for a morphologically rich language - amharic. *In SLTU*.

TACHBELIE, M., ABATE, S. et MENZEL, W. (2011b). Morpheme-based and factored language modeling for Amharic speech recognition. *In* VETULANI, Z., éditeur : *Human Language Technology. Challenges for Computer Science and Linguistics*, volume 6562 de *Lecture Notes in Computer Science*, pages 82–93. Springer.