Corpus based linguistic exploration via forced alignments with a 'light-weight' ASR tool

Jamison Cooper-Leavitt[1], Lori Lamel[1], Annie Rialland[2], Martine Adda-Decker[1,2], Gilles Adda[1]

[1]LIMSI, CNRS, Université Paris-Saclay, France
[2]LPP, CNRS-Paris 3/Sorbonne Nouvelle, France

Abstract

In this work we make use of a baseline ASR system developed for a speech corpus of Embosi (Bantu C25)—a less-resourced language. A first version of this system has been used as a light weight ASR tool to produce forced alignments with the aim of carrying out corpus based linguistic studies. Several linguistic studies of Embosi have identified the deletion of associative morphemes and vowel elision as outstanding issues for further research. We show empirical evidence derived from the Embosi speech corpus that the deletion of these morphemes is not observed equally across all classes, but that there are systematic differences in the occurrence of the associative class morphemes being deleted. We also observe from the corpus that vowel elision interacts with the deletion of these morphemes. We show that with limited language resources, linguistic analysis on less-resourced languages can be accomplished using simple/light-weight models on small speech corpora.

1. Introduction

Embosi (a Bantu language designated as C25) is a less-resourced language spoken in Congo-Brazzaville. There are a limited number of speech tools available for corpus-based empirical studies on Embosi. Yet, it presents several phonological and morphological issues relevant and interesting to linguistic research. In the sense of the limited resources available, we focus this paper on the following question: what automatic speech recognition (ASR) tools can be used to investigate linguistic issues in a small speech corpus of Embosi—a less-resourced language? We present a 'light-weight' ASR model as our tool for an investigation within the scope of a limited Embosi speech corpus. We present some systematic patterns discovered within the corpus through the application of ASR tools that relate to current linguistic topics regarding this language.

Our primary methodology was to develop a tool to perform a series of forced alignment tasks on an Embosi speech corpus (see Adda-Decker et al., 2013) for methods in applying forced alignments to linguistic research). The Embosi data for this study comes from the Breaking the Unwritten Language Barrier (BULB) project (Adda et al., 2016), and the forced alignment task was done using ASR tools from LIMSI's STK speech processing tool kit (Gauvain and Lamel, 2003; Lamel and Gauvain, 2005). Within the scope of this paper we investigate the ASR tool's ability to detect vowel elision and morpheme deletion in the corpus. We use a novel approach as we apply an ASR tool for the purpose of deriving forced alignment segmentations, and we use these segmentations to explore linguistic aspects of the language.

Embosi is spoken by approximately 140,000 number of speakers in the Cuvette Region of Congo. Embosi has very little written material and no official writing convention. It is a typical Bantu language. It has a class system, and it is tonal. Recent studies on Embosi include: several descriptive grammars that describe the Embosi noun class system (Fontaney, 1988; Fontaney, 1989) and its verbal system (Bedrosian, 1996; Amboulou, 1998; Beltzung et al., 2010); a French-Embosi dictionary (Beapami et al., 2000) and an English-Embosi dictionary (Ndongo Ibara, 2012). There have also been several studies on the phonology and phonetics of Embosi (Amborobongui, 2013; Kouarata, 2014; Rialland et al., 2012; Rialland et al., 2015a). These studies have shown that Embosi has several complex phonological processes (e.g. dropping of prefix consonant of some morphemes, and numerous cases of vowel elision) that require further investigation. We present evidence that using ASR tools on a speech corpus of the language provides plenty of examples which show there are systematic and contextual distributions governing the deletion of associative morphemes. Vowel elision also occurs in the context of these morphemes, and in the context in which they are deleted.

1.1. Embosi associative morphemes and vowel elision

The Embosi associative morpheme (also termed as a connective morpheme (Beltzung et al., 2010; Amborobongui, 2013; Kouarata, 2014)) makes explicit the role in which an object or state is associated with another object. This other object is either a person, a thing or an abstract entity. In Embosi the role of this type of morpheme is similar to the French *de*, as in the phrase *l'agrume de Marie* 'the citrus fruit of Marie'. In Embosi, these morphemes are monosyllabic CV words with a single consonant in the onset position and a low vowel with a high tone as the nucleus (i.e. C-*á*).[1] They agree in Noun Class (NC) with the object acting as the head of the phrase. Table 1 shows a list of the Embosi associative morphemes and the agreement with their corresponding NCs.

Example (1) is an example where the associative morpheme *lá* (corresponding to class 5) is not included in the sentence. The omission of the morpheme is considered optional (Kouarata, 2014).

[1]The form of the associative morpheme for class 1 is unique from the other NCs in that it does not have a high tone, but instead has the form C-*a*.

Morph.	Noun Class (NC)
y-a	*1*
y-á	*7*, *8*, *10*
b-á	*2*, *8*, *14*
m-á	*3*, *4*, *6*
l-á	*5*, *11*

Table 1: Embosi associative morphemes

(1) Lidi ilálá íbɔi.
Lydie 5.citrus 5.rotten.PERF
'The citrus of Lydie is rotten.'[2]

In Kouarata's description of the Embosi associative morpheme there are two phonological rules given, which govern the form of the associative morpheme depending on the morphemes context within an utterance. In the first rule, the initial consonant of the associative morpheme is deleted when the following word starts with a consonant (i.e. CV# $>$ V/ _C).[3] Example (2) shows the onset consonant in the associative morpheme *má* being deleted. The vowel is represented in boldface in the given transcription [ɔkándá**á**nɔ] of a phonetic utterance for sentence (2).

(2) ɔkándá má nɔ
3.habit 3.ASC you
'Your habit'

In the second rule, the final vowel (either *-á* or *-a*) is deleted when the following word starts with a vowel (i.e. CV # $>$ C/ _V). This is the case when the following word has more than one syllable and its first syllable does not have a consonant in the onset position (i.e. V.CV). Examples (3) and (4) show the morphemes *má* and *lá* deleting the vowel *-á* when each are followed by a word with a syllabic structure of V.CV. The deleted vowels and remaining consonants are represented in boldface in the phonetic transcription [ɔkándá**mó**obia] for sentence (3) and in the phonetic transcription [ilála**ló**obianɔ] for sentence (4).

(3) ɔkándá má obia
3.habit 3.ASC 1.friend
'habit of someone else'

(4) álá lá obia ya nɔ
5.citrus 5.ASC 1.friend 1.ASC you
'the citrus of your friend'

In each of the phonetic transcriptions the tone of the deleted vowel was added to the vowel of the following word. This rule in which the vowel deletes is known as vowel elision (Rialland et al., 2012; Rialland et al., 2015a; Rialland et al., 2015b).

The vowel elision rules (5) and (6) depend on the contact of vowels across word boundaries. The vowel length (or quantity) is based on long or short values. For cases of long/short vowel contact (see (5)), vowel length, tone, and tone contour are necessary to accurately model the change in vowel length and the change in vowel tone contour. For cases of short/short vowel contact (see (6)), vocalic features of tongue position and tongue height as well as tone are necessary to model the loss of a vowel, and the change in vowel tone.

(5) CV[$long, HL$]#V[$short$].CV $\rightarrow$ CV[$short, H$]#V.CV

(6) CV[H, low]#V[L, mid].CV $\rightarrow$ CV[H, mid]#CV

1.2. The rest of this paper

We discuss the development of the Embosi corpus, a light-weight ASR tool, and the application of these resources to investigate the linguistic issues relating to the associative morphemes and vowel elision. The state of the corpus's development and the ASR tool used for forced alignment are discussed in section 2. We briefly discuss the implications that the deletion of associative morphemes have for Embosi in section 3, and then conclude this paper in section 4.

2. Language resources

There are several elements that are part of the study described in this paper. A corpus of Embosi was developed and segmented at the phone level and at the word level. An ASR tool along with a variant pronunciation dictionary was also developed for the purpose of investigating vowel elision and morpheme deletion in the corpus.

2.1. Embosi BULB Corpus

A small speech corpus of Embosi was derived from two 1-month field trips to Brazzaville-Congo. This work was completed by a native Embosi speaker using the LIG-Aikuma software (Bird and Hanke, 2013; Gauthier et al., 2016). This software is a mobile speech App for Android, which is used to make speech recordings during language-based field work. Built into the LIG-Aikuma software are a series of options. One of these options includes a multitude of culturally specific images (e.g. food items, local wildlife, traditional clothing, etc.). During the field work stage in the development of the corpus, these images were shown to native Embosi speakers in order to elicit spontaneous speech between the field worker and a native Embosi speaker, or between a group of native speakers. These conversations were recorded using the speech App and saved as audio files in a WAV format. These audio files were then segmented into individual utterances for the Embosi corpus. The method used in the preparation of the utterances from the audio files was done with an initial utterance from the native speaker, which was followed by a careful re-speaking from the native speaker, and finally followed by a French translation of the original utterance from the native speaker (Bird et al., 2014). The audio files used in the study of the careful re-speaking were transcribed by native Embosi speakers and are included as accompanying text files corresponding to each utterance's audio file.

In total, the corpus contains 48 hours of speech data (i.e. Bible lectures, verb conjugations and elicited speech), of which a subset of 25 hours have currently been manually

[2]The numerical representations in examples (1) to (4) represent the NC of each noun and associative morpheme. ASC represents the associative morpheme, and PERF represents the perfective verb form.

[3]The # symbol represents a word boundary. For example, 'CV#' represents a CV syllable followed by a word boundary.

transcribed. For this research we have utilized a sub-corpus of 4.5 hours which is in the form of elicited speech (read by 3 male Embosi speakers) in two forms: (i) 1472 utterances, extracted from reference sentences for oral language documentation (Bouquiaux and Thomas, 1976) have been translated and written in Embosi; (ii) 3706 utterances, extracted from an elicitations derived from readings of an Embosi dictionary (Kouarata, 2014). There are a total of 5178 individual utterances each saved as a separate audio files in the corpus.

2.2. A light-weight ASR tool

The ASR based alignment tool uses the STK tools at LIMSI (Gauvain and Lamel, 2003; Lamel and Gauvain, 2005). The primary functionality of the ASR tools used in this work is to perform forced alignment at the lexical level and at the phonetic segment level. We refer to this tool as a 'light-weight' ASR tool. The acoustic model of the tool is a GMM-HMM based monophone model trained from a 'flat start', meaning that there is no prior information given to the system about where the phones occur in the audio signal or even the speech/non-speech separation. Each phone in the system is modeled with a 3-state left-to-right HMM with Gaussian mixture (on average 5 Gaussians per state). The acoustic feature vector is comprised of 12 cepstral coefficients and the fundamental frequency (F0), as well as their first and second delta parameters to capture the dynamic nature of speech (Lamel and Gauvain, 2012). Cepstral mean and variance normalization are carried out for each audio file.

The model is constructed using standard training techniques. In this case 5 iterations of the process of segmenting and model estimation are run, and the model of the last iteration is used to produce the final segmentations of the corpus. Since the data from the Embosi corpus is extremely limited, there is at this time no separate training and test data. However, since the light-weight ASR model is only used for the purpose of forced alignment, and not for any word or phone level recognition testing purposes, we feel that the model is sufficient for the linguistic studies explored in this work where the goal is to identify utterances where morpheme deletion and vowel elision occur. The main motivation for segmenting the data with a monophone model is to minimize the possibility that the acoustic models intrinsically cover some of the phonemic variants we are trying to detect (Adda-Decker and Lamel, 1999).

The tool was used to carry out forced alignment of the 4.5 hour corpus of Embosi, comprised of the transcribed portion of the careful re-speaking utterances. The audio file was the acoustic signal for the forced alignment. The transcription from the native speaker was tokenized into a lexicon and each lexeme was converted into a phonetic pronunciation and stored in a speech dictionary. The lexicon contained 68 phones, and one symbol representing silence.

Embosi vowels include length and tone as suprasegmental features. Tone is traditionally not represented, however, as a vocalic feature for Bantu languages. Instead it is treated as a syllabic feature (Hooper, 1976). In other words, phonological and morphological analyses where the tone bearing unit is considered to be a syllabic unit instead of a vocalic are considered to be more accurate in representing the complex tonal morphology and phonology of these languages. With the limited amount of training data, it is not possible to accurately represent this with the alignment tool. Pitch (measured in *Hz* by autocorrelation) is implicitly captured in the acoustic model used by the ASR tool based on the the tonal information in the acoustic signal.

Explicitly representing vowels with features of tone and length in this manner greatly increased the number of phones used in the model (from 31[4] to 68). However, even with the small set of utterances in the corpus, the model was able to get enough representation of each phone for the ASR tool to accurately force align segments to the speech signal. A more comprehensive solution to more accurately treat tone in the model still stands out as an issue for further development of the ASR tool.

2.3. Variant pronunciation dictionary

A dictionary of speech variants was created for the ASR tool. For each text file containing the phonemic transcriptions of utterances that included an associative morpheme, an underscore character (i.e. '_') was added to the sentence to concatenate the associative morpheme with the word immediately to its right (the possessor in the associative phrase) in the utterance. For the purpose of the ASR tool, this concatenated string was treated as a single lexical item, even though it was literally a string of 2 words. This *new* lexical item was added to the dictionary. Several variant pronunciations were generated for each of these concatenated items in two contexts: (a) the concatenated left edge morpheme was deleted, and only the right edge word was used to generate phonetic symbols to represent pronunciation variants; (b) the entire string of concatenated words/morphemes were used to generate phonetic symbols to represent pronunciation variants. Further considerations were also made to represent possible phonetic variations due to vowel elision internal to the concatenated string of morphemes, and to represent any possible phonetic variations due to vowel elision at the right and left edges of the concatenated strings.

The following possible conditions were represented as variants in the dictionary: (a) a condition in which no morpheme deletion occurred, but vowel elision did occur between the concatenated words and morphemes; (b) a condition in which morpheme deletion and vowel elision occurred between the immediate word to the left and the word to the right of the concatenated string. In this last condition, the deleted word was assumed to not be phonetically included for the purpose of vowel elision.

3. Frequency of morpheme deletion and vowel elision

The ASR tool predicted the deletion of the associative morphemes in many of the utterances in the corpus. The frequency of deletion of morphemes, along with the occurrence of vowel elision are shown in table 2. An orthographically truncated form (i.e. *y'*, *l'*, *b'* and *m'*) was used

[4]There are 24 consonant and 7 vowel phonemes in Embosi (Bedrosian, 1996).

by some of the native transcribers for utterances in which they considered vowel elision had occurred. The frequency results for these truncated morphemes is included in table 2. For these truncated forms, the transcribers had predetermined that vowel elision occurred between the associative morpheme and the following word in the utterance. Therefore the ASR tool would not be able to detect vowel elision in these cases, because the variant dictionary used for the model would not include the vowel in the first place. The values for the truncated morphemes with vowel elision were omitted, since in all these cases vowel elision was already predetermined by the transcribers to have occurred.

The corpus provides plenty of examples which support arguments claiming morpheme deletion and vowel elision affects the phonetic representation of Embosi (Rialland et al., 2012; Amborobongui, 2013; Kouarata, 2014). The morpheme *ya* (i.e. /ja/) only corresponds to NC 1, and in 75.9% of instances where *ya* was transcribed by native speakers, the ASR alignment tool determined that *ya* was deleted. We observe that this morpheme has a high frequency of deletion compared to the other associative morphemes. The associative morpheme with the next highest frequency of deletion is *yá* (i.e. /já/), where the ASR alignment tool determined it was deleted in 51.2% of instances. The truncated morpheme *y'* was deleted in 39% of instances, but it is unclear to which form (*ya* or *yá*) and noun class it corresponds. All other associative morphemes have a frequency of deletion (according to the ASR alignment tool) of 27.3% or less. This compelling evidence derived from the tool suggests that the morpheme *ya* is treated differently by speakers than the other associative morphemes.

Native speakers may be aware and often anticipate missing elements in an utterance (Nooteboom, 2001). In our case native language experts trained in language transcription chose to transcribe the symbol *l'* for a CV morpheme when they considered the vowel in that morpheme to be omitted. This also occurred for the other associative morphemes. The deliberate marking of vowel elision on the part of the native transcribers is evidence that they are often aware that vowel elision occurs.

Another observation is that native transcribers may mentally restore elements which are not present in the acoustic signal of the utterance. From the transcriptions provided by the native speakers, there exists many instances where the associative morpheme was transcribed, but the ASR tool detected that it was deleted. A similar case can be made about vowel elision, where the ASR tool detected vowel segments being elided in the acoustic signal of many utterances, but the native speakers still transcribed them as being present. Native speakers may often be perceptually inclined to mentally repair an utterance when the acoustic signal is missing or it has incorrect phonetic information. Some ASR tools have higher level functioning in learning to repair such a signal, but this capability is not developed in the ASR tool used on this corpus.

Using light-weight ASR tools holds a significant advantage when being applied to linguistic research in an LRL environment. A challenge for this type of research is in gathering enough language/speech data to train and test more sophisticated models for ASR. Typical sources of language data gathering are not readily available (i.e. crowd sourcing, phone banks, etc.). In the case of the Embosi as part of the BULB project, extensive field research is needed, which also includes a significant degree of post-development of the data to be suitably used in the corpus. Furthermore, less-resource languages may also contain unique phonological and morphological properties. Monophone models for segment alignment purposes are often more useful for linguistic investigations. If more context is used by the model, then it does not make use of the dictionary. The model intrinsically makes phone variations that are not explicitly done with the variant dictionary. Therefore, it is more difficult to track these variations which may be of interest to linguistic analyses.

4. Conclusion

In this paper we have applied forced alignment to investigate the linguistic topics of the deletion of associative morphemes and vowel elision. We used a light-weight ASR tool that is a GMM-HMM based monophone model. This was advantageous to us in two respects. First, it allowed us to more accurately observe and count the variants for vowel elision and deletion that were included in the variant dictionary. The reliance on this dictionary allowed us to count the occurrences of when the ASR tool detected these linguistic issues. Second, due to Embosi being a less-resourced language, the speech corpus was extremely limited and insufficient for a more sophisticated and deeper model. We were successful in determining systematic occurrences of the deletion of the associative morpheme that are interesting to linguistic research.

5. Acknowledgments

This work was partly funded by the French ANR and the German DFG project BULB under grant ANR-14-CE35-0002, and the Labex EFL program under grant ANR-10-LABX-0083.

6. References

Adda, Gilles, Sebastian Stuker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noel Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian, 2016. Breaking the unwritten language barrier: The BULB project. *Procedia Computer Science*, 81:8–14.

Adda-Decker, Martine, Giles Adda, and Lori Lamel, 2013. *Méthodes et Outils pour l'Analyse Phonétique des Grands Corpus Oraux*, chapter Systèmes de Transcription comme Instruments. Cachan, France: Hermès Science, pages 159–202.

Adda-Decker, Martine and Lori Lamel, 1999. Pronunciation variants across system configuration, language and speaking style. *Speech Communication*, 29:83–89.

Amborobongui, Martial Embanga, 2013. *Processus segmentaux et tonals en Mbondzi - (variété de la language ɛmbɔ́sí C25))*. Ph.D. thesis, Universite Sorbonne Nouvelle - Paris 3.

	N	n_{del}	n_{del+el}	n_{el}
ya	349	265 (75.9%)	4 (1.5%)	14 (4%)
yá	449	230 (51.2%)	6 (2.6%)	25 (5.6%)
y’	187	73 (39%)	31 (42.5%)	—
bá	172	47 (27.3%)	0	16 (9.3%)
b’	111	23 (20.7%)	5 (21.7%)	—
má	288	95 (33%)	1 (1.1%)	22 (7.6%)
m’	565	80 (14.2%)	36 (45%)	—
lá	352	15 (4.3%)	0	20 (5.7%)
l’	531	116 (21.8%)	32 (27.6%)	—

Table 2: Number and frequency of associative morphemes in the Embosi corpus. N is the total number of each morpheme in all utterances of the corpus. n_{del} is the number of occurrences each morpheme is deleted. n_{del+el} is the number of occurrences in which each morpheme is deleted and vowel elision occurs between the possessed object and the possessor. n_{el} is the number of occurrences that the associative morpheme undergoes vowel elision and is *not* deleted.

Amboulou, Clestin, 1998. *Le Mbochi, langue bantou du Congo-Brazzaville : tude descriptive*. Ph.D. thesis, Institut national des langues et civilisations orientales (Paris).

Beapami, Roch Paulin, Ruth Chatfield, Guy Kouarata, and Andrea Waldschmidt, 2000. *Dictionnaire Mbochi - Français*. Brazzaville: SIL-Congo.

Bedrosian, Patricia, 1996. The Mboshi noun class system. *Journal of West African Languages*, 26(1):27–47.

Beltzung, Jean-Marc, Annie Rialland, and Martial Embanga Aborobongui, 2010. Les relatives possessives en ɛmbɔ́sí. *ZAS Papers in Linguistics*, 53:7–37.

Bird, Steven and Florian Hanke, 2013. Aikuma-home. http://www.aikuma.org/. Accessed: 2017-10-18.

Bird, Steven, Florian Hanke, Oliver Adams, and Haejoong Lee, 2014. Aikuma: A mobile app for collaborative language documentation. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Baltimore, Maryland, USA: Association for Computational Linguistics.

Bouquiaux, Luc and Jacqueline Thomas, 1976. *Enquête et description des langues à tradition orale.Tome II: Approche linguistique (questionnaires grammaticaux et phrases).*, volume II of *SELAF*. Peeters Publishers.

Fontaney, Louise, 1988. *Pholia*, chapter Mboshi steps towards a grammar Part 1. Université Lumière Lyon 2, pages 87–167.

Fontaney, Louise, 1989. *Pholia*, chapter Mboshi steps towards a grammar Part 2. Université Lumière Lyon 2, pages 71–131.

Gauthier, Elodie, David Blachon, Laurent Besacier, Guy-Noel Kouarata, Martine Adda-Decker, Annie Rialland, Gilles Adda, and Grégoire Bachman, 2016. LIG-Aikuma: a Mobile App to Collect Parallel Speech for Under-Resourced Language Studies. In *Interspeech 2016 (short demo paper)*. San-Francisco, France.

Gauvain, Jean-Luc and Lori Lamel, 2003. *Pattern Recognition in Speech and Language Processing*, chapter Large vocabulary speech recognition based on statistical methods. CRC Press, pages 149–189.

Hooper, Joan, 1976. *An Introduction to Natural Generative Phonology*. New York: Academic Press.

Kouarata, Guy Noel, 2014. *Variations de Formes Dans la Langue Mbochi (Bantu C25)*. Ph.D. thesis, Université Lumière Lyon 2.

Lamel, Lori and Jean-Luc Gauvain, 2005. *The Oxford Handbook of Computational Linguistics*, chapter Speech recognition. Oxford: Oxford University Press, pages 305–322.

Lamel, Lori and Jean-Luc Gauvain, 2012. *The Oxford Handbook of Computational Linguistics*, chapter Speech Recognition. Oxford University Press.

Ndongo Ibara, Yvon Pierre, 2012. *Embosi-English Dictionary*. Peter Lang.

Nooteboom, Sieb, 2001. Different sources of lexical bias and overt self-corrections. In *ISCA Archive*.

Rialland, Annie, Martial Embanga Amborobongui, Martine Adda-Decker, and Lori; Lamel, 2012. Mbochi: corpus oral, traitement automatique et exploration phonologique. In *Traitement Automatique des Languages Africaines*.

Rialland, Annie, Martial Embanga Amborobongui, Martine Adda-Decker, and Lori; Lamel, 2015a. Phonologie et traitement automatique de la parole: le cas de l’embosi (bantu c25). In *The American Conference of African Linguistics*.

Rialland, Annie, Martial Embanga Aborobongui, Martine Adda-Decker, and Lori Lamel, 2015b. Dropping of the class-prefix consonant, vowel elision and automatic phonological mining in embosi (Bantu C 25). In *Selected Proceedings of the 44th Annual Conference on African Linguistics*, Cascadilla Proceedings http://www.lingref.com/cpp/acal/44/index.html. Georgetown University: Ruth Kramer, Elizabeth C. Zsiga, and One Tlale Boyer.