A Corpus for Amharic-English Speech Translation: The Case of Tourism Domain

Michael Melese Woldeyohannis[1](✉), Laurent Besacier[2], and Million Meshesha[1]

[1] Addis Ababa University, Addis Ababa, Ethiopia
{michael.melesele,million.meshesha}@aau.edu.et
[2] LIG Laboratory, UJF, BP53, 38041 Grenoble Cedex 9, France
laurent.besacier@imag.fr

Abstract. Speech translation research for the major languages like English, Japanese and Spanish has been conducted since the 1980's. But no attempt were made in speech translation to/from the under-resourced language like Amharic. These activities suffered from the lack of Amharic speech and Amharic-English text corpus suited for the development of speech translation between the two languages. In this paper, therefore, an attempt has been made to collect, translate and record speech data from resourced language (English) to under-resourced language (Amharic) taking a Basic Traveler Expression Corpus (BTEC) as domain. Since there is no any Amharic text and speech corpus readily available for speech translation purposes, first, 7.43 h of Amharic read-speech has been prepared from 8,112 sentences, and second, 19,972 parallel Amharic-English corpus has been prepared taking tourism as an application domain. The Amharic speech data is recorded using smartphone based application tool, LIG-Aikuma under a normal working environment. With the availability of such standard speech and text corpus, researcher will find a ground to further explore speech translation to/from under resourced languages.

Keywords: Amharic speech corpus · Basic traveller expression corpus
Amharic-English corpus

1 Introduction

Speech is the most natural form of communication in an increasingly globalized world economy, national security and humanitarian service [1]. Alongside this, computers with the ability to understand speech spoken in different languages greatly contribute for the development of man-machine interfaces [2]. This can be extended through different digital platforms such as radio, mobile, TV, CD and others.

1.1 Speech Translation

Research in speech translation for technological supported major languages like English, European languages (like French and Spanish) and Asian languages

F. Mekuria et al. (Eds.): ICT4DA 2017, LNICST 244, pp. 129–139, 2018.
https://doi.org/10.1007/978-3-319-95153-9_12

(like Japanese and Chinese) has been conducted since the 1983s by NEC Corporation [3]. The advancement of speech translation captivates the communication between people who do not share the same language.

The state-of-the-art speech translation system can be seen as a cascade of three major components [4]; Automatic Speech Recognition (ASR), Statistical Machine Translation (SMT) and Text-To-Speech (TTS) synthesis. ASR is the process by which a machine identifies spoken words, by means of talking to computer, and having it correctly understand what the speaker is saying [5]. Beside ASR, SMT deals with mapping of sentences from one source language into another target language using a model projected from parallel corpora automatically [6]. Finally, TTS system converts the text data into synthesized speech [7].

As one main component of speech translation, Amharic ASR started in 2001. As discussed in Melese et al. [8], several attempts have been made for Amharic ASR using different methods and techniques to solve a different issues encountered during speech recognition. Besides ASR, preliminary experiments on English-Amharic machine translation has been conducted using phonemic transcription on the Amharic corpus and encouraging result were found [9].

As a last component of speech translation, a number of TTS research have been attempted using a number of techniques and methods as discussed by Anberbir [10]. Among these, concatenative, cepstral, formant and a syllable based speech synthesizers were the main methods and techniques applied.

However, all the above researches conducted using different domain data, methods and techniques. Beside this, dataset used in the above research not available, in addition to the difficulty to evaluate the advancement of research in speech technology for local languages. As a result, these activities suffer from the lack of unavailability of Amharic-English text and Amharic speech corpus. Thus, collecting, preparing a text and speech corpus for the development of Amharic-English speech translation system alleviates the problem of data inadequecy.

1.2 The Need for Speech Translation in Tourism Domain

Tourism involves the direct contact between people and cultures, becoming pleasing sustainable economic development and serving as an alternative source of foreign exchange for the country like Ethiopia [11]. According to the official site of the Ethiopian Embassy in the USA [12], Ethiopia has much to offer for international tourists. The historic root comprises of ancient and medieval cities including world heritages, which is registered as Ethiopian tourist attraction by UNESCO.

Moreover, The 2015 United Nations World Tourism [13] and the World Bank[1] report indicate that, a total of 864,000 non-resident tourists come to Ethiopia to visit different locations; out of more than 1 billion international tourists. Since the year 2010 until 2015, the number of tourist increase every year on average by 13.9%.

[1] http://data.worldbank.org/indicator/ST.INT.ARVL.

As a result, Walta Information Center[2], citing Ethiopia Ministry of Culture and Tourism, noted that the country has secured about 3.3 billion dollar in the 2016/17 fiscal year from more than 886,000 international tourists. The revenue was collected from various tourist attraction sites across the country, majority of the tourists were from USA, England, Germany, France and Italy and they don't speak any of the Ethiopian languages. These tourists speak foreign languages hindering them to communicate in local languages with tourist guides.

Hence, language barrier is a major problem for today's global communication. Beside this, tourist express their idea using different languages, majority of the tourist can speak and communicate in English to exchange information about tourist attractions. As a result, they look for an alternate option that lets them to communicate with the environment. Thus, speech translation system is one of the best technology used to fill the communication gap between the people who speak different languages.

However, under-resourced languages such as Amharic, suffer from having a text and speech corpus to support speech translation technology. Therefore, the main aim of this paper is to develop Amharic speech corpus and construct Amharic-English parallel text corpus for speech translation in the tourism domain taking basic traveler expression as a domain. Accordingly, moving one step further by collecting resource for under-resourced language helps in overcoming language barriers in today's global communication.

2 Amharic Language

Ethiopia has 89 different languages which are registered in the country with up to 200 different spoken dialects [14]. Among these languages, Amharic is the official working language of government of Ethiopia, and some regional states, such as Addis Ababa, Amhara, Diredawa and Southern Nations, Nationalities and People (SNNP). As Semitic language Amharic is a derived from Ge'ez with the second largest speaker in the world next to Arabic. The name Amharic (አማርኛ-amarəñña) comes from the district of Amhara (አማራ) in northern Ethiopia, which is thought to be the historic center of the language [15]. Moreover, the language Amharic is used in governmental administration, public media and mass communication (television, radio, literature, entertainment, etc.), and national commerce.

According to Language of the world [15], the number of Amharic speaker is more than 25 million with up to 22 million native speakers. The majority of Amharic speakers found in Ethiopia even though there are also speakers in a number of other countries, particularly Italy, Canada, the USA and Sweden.

In the following section a review of Amharic writing system and its phonetics is given in view of preparing speech and text corpus for the development of Amharic-English speech translation.

[2] http://www.waltainfo.com/news/detail/32016.

2.1 Amharic Writing System

Amharic characters represent a consonant vowel (CV) sequence and the basic shape of each character is determined by the consonant, which is modified for the vowel. Unlike other Semitic languages, such as Arabic and Hebrew, modern Amharic script has inherited its writing system from Ge'ez (ግእዝ) /gə'əzə/·. Amharic script uses a grapheme based writing system called fidel /fidälə/ which is written and read from left to right being the classical and ecclesiastical language of Ethiopia.

Amharic writing system is categorized into four distinct categories consisting of 276 different symbols [16]; 231 core characters, 20 labiovelar symbols, 18 labialized consonants and 7 labiodental characters. The first category possess 33 primary characters each representing a consonant having 7 orders (ə/አ/, u/ኡ/,i/ኢ/, a/ኣ/, e/ኤ/, ʔ/እ/, o/ኦ/) in form to indicate the vowel which follows the consonant to represent CV syllables.

Likewise, labiovelar symbols contains 4 (ቈ/k'/, ኈ/h/, ኰ/k/ and ጐ/g/) characters with 5 orders (^{w}ə, ^{w}i, ^{w}a, ^{w}e and ^{w}) that generates 20 distinct symbols. Similarly, labiodental category possess 1 character (ቨ/v/) with 7 order like core characters and only appears in modern loan words borrowed from foreign languages like ቪዛ/viza/. There are also labialized 18 characters; for instance, ሏ/l^{w}a/, ሟ/m^{w}a /, ሯ/r^{w}a / and ሷ/s^{w}a / that are used in Amharic writing system.

In Amharic writing, all the 276 distinct symbols are indispensable due to their distinct orthographic representation. However, for cascading components of speech translation, we mainly deal with distinct sound in speech recognition and orthographic representation for machine translation.

Thus, among the given character set, different graphemes that generate the same sound has normalized to minimize sounds and words modelled in speech recognition and machine translation respectively. Table 1 presents variants of Amharic graphemes that has been normalized into common graphemes.

Table 1. Amharic characters normalization (adopted and modified from [8])

Variants	Count	Normalized
ሀ, ሐ, ኀ and ኸ	4	ሀ
አ and ዐ	2	አ
ሠ and	2	ሰ
ፀ and ጸ	2	ፀ

Among the given 33 core character set, graphemes with multiple variants have to be normalized into their orders to generate equivalent graphemes as shown in Table 1. The selection of graphemes is made based on the usage of most characters frequency in Amharic writing system. Thus, as a result of normalizing the seven orders of (ሀ, ሐ, ኀ, ኸ) to ሀ/h/, (አ, ዐ) to አ/ʔ/, (ሠ, ሰ) to ሰ/s/, (ፀ, ጸ) to ፀ/ts'/ there is a great reduction in one-to-many, many-to-one and many-to-many model of the cascading component of speech translation.

2.2 Amharic Phonetics

The Amharic phoneme inventory is comprised of 38 phones, 31 consonants and 7 vowels which handles a complete set of sound for Amharic language [17]. Consonants are classified into three categories; manner of articulation, place of articulation, and voicing method [16]. The manner of articulation refers to the interaction of speech organs such as the tongue, lips, and palate when making a speech sound. Similarly, place of articulation indicate where the air flow is blocked in the mouth in order to create sound. Voicing determine whether the sound in query is pronounced while the vocal folds are vibrating or not. Table 2 depicts a complete set of Amharic consonant with respect to their manner of articulation, voicing, and place of articulation.

Table 2. Amharic consonants categories.

Manner of articulation	Voicing	Place of articulation				
		labial	dental	palatal	velar	glottal
Stop	voiceless	ፕ	ት	ች	ክ	እ
	voiced	ብ	ድ	ጅ	ግ	
	glottalized	ጵ	ጥ	ጭ	ቅ	
	rounded				ኵ ጕ ቍ	
Fricatives	voiceless	ፍ	ስ	ሽ		ህ
	voiced		ዝ	ዥ		
	glottalized		ጽ			
	rounded					ኍ
Nasals	Voiced	ም	ን	ኝ		
Liquids			ል ር			
Semi-vowel		ው			ይ	

Based on the manner of articulation, Amharic sound consist of stops, fricatives, nasals, liquids, and semi-vowels sound whereas taking place of articulation, labial, dental, palatal, velar and glottal sound. Voicing consist of voiced, voiceless, glottalized and rounded type sound.

Beside the consonant, Amharic has a seven-term vowel system (ኧ[ə], ኡ[u],ኢ[i], ኣ[a], ኤ[e], እ[ʔ] and ኦ[o]) that are categorized as rounded (ኡ[u] and ኦ[o]) and unrounded (ኧ[ə], ኢ[i], ኣ[a], ኤ[e] and እ[ʔ]). .

Likewise, Amharic vowels are categorized according to the place of articulation which includes horizontal movement of tongue such as front, center and back and vertical movement of tongue such as high, middle and low as depicted in Fig. 1. The place of articulation of the Amharic vowel ኧ[ə] is middle and center, ኡ[u] back and high, ኢ[i] front and high, ኣ[a] center and low, ኤ[e] front and middle, እ[ʔ] center and high while ኦ[o] is back and middle based on the movement of the tongue in mouth.

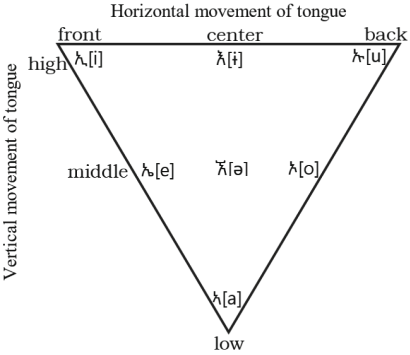

Fig. 1. Category of Amharic vowels in terms of articulation

3 Corpus Preparation

Though speech and text corpora are one of the fundamental resources for any speech translation system; Collecting and preparing standardized and annotated corpora is a challenging and expensive task [14]. This is especially true when working with under-resourced languages like Amharic; under-resourced language require innovative data collection methodologies since there is a lack of language resources that suffers from devising text and speech corpora in digital format. Beside these, due to inaccessibility of standardized digital corpora in tourism and other domains, a text corpus is acquired from resourced and technologically supported languages. Accordingly, parallel English-Arabic text corpus was acquired from BTEC 2009 which is made available through International Workshop on Spoken Language Translation (IWSLT) [18]. In this corpus preparation, currency, time, date, numbers and location have been translated as it is. Moreover, one English sentence is translated in to multiple equivalent Amharic sentence taking the most common expression used in Amharic document as shown in Table 3.

English sentence "I have to catch a plane for Los Angeles early tomorrow morning" is translated in to four different sentence due to equivalence of meaning "**የሚበረውን**" with "**የሚሄደውን**" and "**መያዝ**" with "**ላይ መሳፈር**" as underlined in Table 3 without considering grapheme variation.

Similarly un-normalized data leads to possible multiple translation for Amharic equivalent even with a single graphemes change. If we take English sentence "I'll take my son to the zoo" and the equivalent translation "**ልጄን ዛሬ ወደ እንስሳት መንከባከቢያ ዕወሥደዋለሁ**", there are $2^5(32)$ possible way of writing for a single English sentence as a result of using the graphemes **እ** or **ዕ** to **እ**, **ሰ** or **ሥ** to **ሰ** and **ሣ** to **ሳ** to **ሳ** in the word "**እንስሳት**" and "**ዕወሥደዋለሁ**". Thus, normalizing these graphemes to common writing system results to one translation equivalent than 32.

This corpus consists of 28,084 sentences in basic traveller expression domain of which 19,972 is to be used for machine translation and the remaining 8,112

Table 3. One-to-many translation of English to Amharic sentence.

English Sentence	Possible Amharic translation
I have to catch a plane for Los Angeles early tomorrow morning.	ነገ ጠዋት ወደ ሎስአንጀለስ የሚሄደውን አውሮፕላን መያዝ አለብኝ ።
	ነገ ጠዋት ወደ ሎስአንጀለስ የሚበረውን አውሮፕላን መያዝ አለብኝ ።
	ነገ ጠዋት ወደ ሎስአንጀለስ የሚሄደውን አውሮፕላን ላይ መሳፈር አለብኝ ።
	ነገ ጠዋት ወደ ሎስአንጀለስ የሚበረውን አውሮፕላን ላይ መሳፈር አለብኝ ።
Can I cook by myself ?	እራሴ ምግብ ማብሰል እችላለሁ ?
	እራሴ ምግብ ማዘጋጀት እችላለሁ ?
	እኔ ራሴ ምግብ ማብሰል እችላለሁ ?
	እኔ ራሴ ምግብ ማዘጋጀት እችላለሁ ?
Which floor is the kitchen appliances department ?	የኩሽና እቃዎች ያሉበት ስንተኛው ወለል ላይ ነው ?
	የኩሽና እቃዎች ያሉበት የትኛው ወለል ላይ ነው ?
	የኩሽና እቃዎች ያሉት የትኛው ወለል ላይ ነው ?
	የኩሽና እቃዎች ያሉት ስንተኛው ወለል ላይ ነው ?

used for Amharic speech recognition. Section 3.1 discuss the Amharic-English machine translation corpus while Sect. 3.2 discuss the Amharic speech corpus as part of cascading components.

3.1 Amharic-English Text Corpus

Parallel text corpus is required for designing machine translation system. Accordingly, from a parallel English-Arabic corpus, the English part is translated to Amharic to prepare a parallel Amharic-English BTEC corpus using a bilingual speaker linguistic expert. After translation, the Amharic data have been transcribed into Unicode; then, to keep the dataset consistent, the text corpus has been further preprocessed, such as typing errors are corrected, abbreviations have been expanded, numbers have been textually transcribed and concatenated words have been separated. Table 4 presents the distribution of parallel Amharic-English sentence prepared for machine translation.

Table 4. Amharic-English machine translation corpus distribution

	Sentence	Token	Type	
			Unnormalized	Normalized
Amharic	19,972	109,528	20,210	19,014
English	19,972	152,918	8,389	7,357

In addition to this, the translated text corpus collected has been normalized to their orders for consistency of Amharic and English sentences in the data[3].

[3] Normalization in Amharic refers to converting the common graphemes of writing to one while in English it represent lowering the case.

This results in vocabulary size reduction. However, the reduction is minimal due to the use of domain specific text corpus. During translation of English to Amharic, multiple words in English are translated into limited Amharic words, which decrease the number of words in Amharic as compared to the size of words in English as reflected in Table 3. This is due to the morphological richness of Amharic than English.

The Amharic corpus has a word length ranging from 1 to 40 (average of 5 to 6) while, the English corpus consists of 1 to 63 (average of 7 to 8) words per sentence. Figure 2 presents distribution of Amharic and English sentence length against the frequency. Figure 2 presents distribution of Amharic and English sentences interms of frequency of occurence of sentence length (words per sentence)

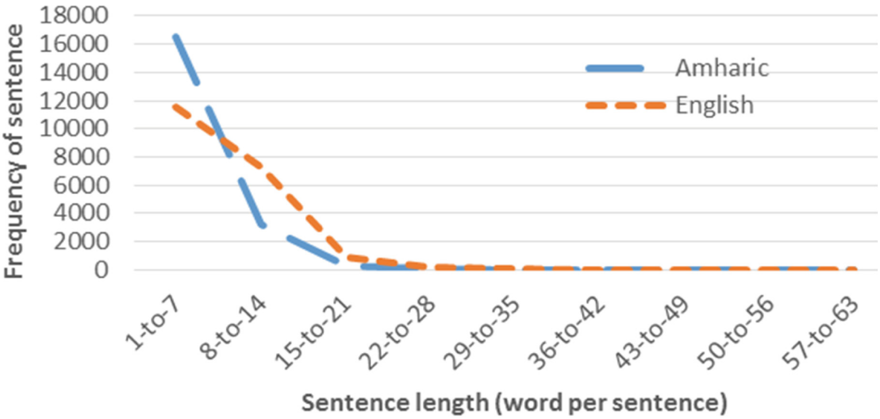

Fig. 2. Distribution of length vs frequency of Amharic-English sentences

Figure 2 depicts that, around 82% of Amharic and 58% of English corpus tend to fall below word length of 8 per sentence, which is below the average due to the domain restriction for English language. Whereas for Amharic, there is no standard average sentence length to the knowledge of the researcher.

Beside sentence length, the Amharic machine translation corpus has 0.71% of Amharic words occurred more than 100 times and 62.7% of Amharic words occurred only once. Similarly, the English data has 44.2% words with frequency of one and 2.9% frequent words more than 100 time occurence.

In general, fact Fig. 2 we found, as the frequency of the sentence increases the sentence length decreases for both languages and the rate of decrease is at an increasing rate for Amharic and at a decreasing rate for English.

3.2 Amharic Speech Corpus

As discussed by Davel [19], traditional methods of data collection is not convenient, consistent and usually time taking particularly for collecting large amount of data compared to handheld device like mobile phones. Beside this, mobile and handheld devices are becoming increasingly available and sharply decreasing cost even for the developing country to collect speech data. Thus, for our typically under-resourced conditions, we opted to incorporate a smartphone based application called LIG-Aikuma [20] to facilitate the speech data collection process under normal office environment. Aikuma does not rely on internet connection to perform audio data collection, but does require that text prompts be loaded on the device manually. The main reason for using LIG-Aikuma is due to its ease of use and an open-source tool that runs on Android platform.

The researcher explained the purpose of the research and instructed to start the recording when they are ready to read in the absence of the researcher. The speech data is collected from eight native Amharic speakers (4 male and 4 female) with different age range for a total of 8112 sentences. Table 5 presents the distribution of the utterance per age and gender of the speaker. The speakers read the Amharic parts of Amharic-English aligned sentence with the possibility to record over again the sentence anytime they mispronounced the sentence. The speech data have been recorded with a length ranging from 1 to 28 word length (average of 4 to 5 words) per sentence. A total of 7.43 h read speech corpus ranging from 1,020 ms to 14,633 ms with an average speech time of 3,297 ms was collected. The distribution of speech length across sentence is presented as depicted in Fig. 3.

Table 5. Age and gender distribution of Amharic speech corpus.

	Age and Gender			
	Male		Female	
	18–30	31–50	18–30	31–50
Number of utterance	1,000	1,112	1,000	1,000
	1,000	1,000	1,000	1,000
Total	2,000	2,112	2,000	2,000

As we can see from Fig. 3, 98.54% of the speech data fall below 7 s. Moreover, the speech data corpus consist of 8112 sentences (37,288 tokens) of 4,168 types. In addition to this, 1.2% of Amharic words occurred more than 100 times.

On the contrary, 42.4% of Amharic words occurred once from ASR data. BTEC developed as a wide-coverage, consistent corpus containing basic travel expressions in technologically supported language, for the purpose of providing basic data for the development of high quality speech translation systems.

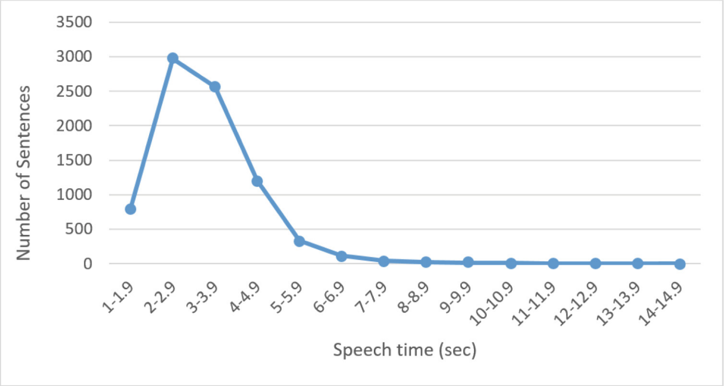

Fig. 3. Distribution of Amharic speech dataset

4 Concluding Remarks

Under-resourced language, like Amharic needs to be supported by technology, like speech translation. To this end, standardized corpus preparation is a must to use it as a test bed to control and evaluate the progress of the research in the area of speech translation. Attempts in this area from under-resourced languages like Amharic, is particularly, not yet started. To ease speech translation from under-resourced language (Amharic) to resourced language (English) text and speech corpus have been developed by translating traveller expression corpus.

Accordingly we have constructed a 19,972 Amharic-English parallel corpus for machine translation and a 7.43 h read-speech corpus from 8112 sentences in tourism domain from widely covered, consistent and technologically supported English language. Thus, the corpus we developed will be used for Amharic-English speech translation by means of cascading components, which is our immediate research direction.

References

1. Gao, Y., Gu, L., Zhou, B.: Speech-to-Speech Translation. World Scientific, Singapore (2007)
2. Honda, M.: Human speech production mechanisms, NTT Commun. Sci. Lab. **1**(2), 24–29 (2003)
3. Karematsu, A., Morimoto, T.: Automatic Speech Translation: Fundamental Technology for Future Cross-Language Communication, vol. 11. Gordon and Breach Publishers, Philadelphia (1996)
4. Gao, Y., Gu, L., Zhou, B., Sarikaya, R., Afify, M., Kuo, H.-K., Zhu, W., Deng, Y., Prosser, C., Zhang, W., Besacier, L.: IBM mastor: multilingual automatic speech-to-speech translator. In: Proceedings of 2006 IEEE International Conference Acoustics Speech Signal Processing, vol. 5, p. 5, December 2006
5. Jurafsky, D., Martin, J.H.: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second edn. Pearson

6. Philipp, K.: Statistical Machine Translation. Cambridge University Press, Cambridge (2009)
7. Lemmetty, S.: Review of Speech Synthesis Technology. Helsinki University of Technology, Helsinki (1999)
8. Woldeyohannis, M.M., Besacier, L., Meshesha, M.: Amharic speech recognition for speech translation. In: JEP-TALN-RECITAL 2016. Paris, France, vol. 11, p. 114 (2016)
9. Teshome, M.G., Besacier, L., Taye, G., Teferi, D.: Phoneme-based English-Amharic statistical machine translation. Presented at the AFRICON, 2015, Addis Ababa, Ethiopia, pp. 1–5 (2015)
10. Anberbir, T., Takara, T.: Development of an Amharic text-to-speech system using cepstral method. In: Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT 2009), Athens, Greece, pp. 46-52 (2009)
11. Ethiopia: A Preferred Location for Foreign Direct Investment in Africa, Ethiopian Investment Commision (2014). http://www.investethiopia.gov.et/images/pdf/Investment_Brochure_to_Ethiopia.pdf
12. Ethiopian Embassy: Investing in ethiopia: Briefing for tour operators. http://www.ethiopianembassy.org/PDF/InvestingTourism.pdf. Accessed 10 Aug 2017
13. UNWTO: World Tourism Organization annual report 2015. Technical report, United Nation, Madrid, Spain (2016)
14. Simons, G.F., Fennig, C.D.: Ethnologue: Languages of the World. SIL, Dallas (2017)
15. Teklehaimanot, T.: Ethiopian Languages. http://www.ethiopiantreasures.co.uk/pages/language.htm. Accessed 10 Aug 2017
16. Abate, S.T., Menzel, W., Tafila, B.: An Amharic speech corpus for large vocabulary continuous speech recognition. In: Proceedings of the XVth International Conference of Ethiopian Studies, Hamburg, Germany (2005)
17. Yimam, B.: Yeamarigna sewasew (Amharic version). Addis Ababa, Ethiopia, EMPDA (1986)
18. Fondazione Bruno Kessler: International Workshop on Spoken Language Translation, Paris, France, pp. 2–3 (2010)
19. Davel, M.H., Badenhorst, J., Basson, W.D., De Wet, F., Barnard, E., De Waal, A., De Vries, N.J.: A smartphone-based ASR data collection tool for under-resourced languages, vol. 56, pp. 119–131 (2014)
20. Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., Rialland, A.: Parallel speech collection for under-resourced language studies using the Lig-Aikuma mobile device app. In: SLTU (2016)