A Parallel Corpora for bi-directional Neural Machine Translation for Low Resourced Ethiopian Languages

Conference Paper · September 2021

DOI: 10.1109/ICT4DA53266.2021.9672230

CITATIONS
8

READS
394

3 authors:

Atnafu Lambebo Tonja
Centro de Investigación en Computación del IPN
30 PUBLICATIONS **28** CITATIONS
SEE PROFILE

Michael Melese Woldeyohannis
Addis Ababa University
27 PUBLICATIONS **121** CITATIONS
SEE PROFILE

Mesay Gemeda Yigezu
Interino del Centro de Investigación en Computación Instituto Politécnico Nacional
7 PUBLICATIONS **15** CITATIONS
SEE PROFILE

A Parallel Corpora for bi-directional Neural Machine Translation for Low Resourced Ethiopian Languages

Atnafu Lambebo Tonja
Department of Information Technology
Wolaita Sodo University
Wolaita Sodo, Ethiopia
atnafu.lambebo@wsu.edu.et

Michael Melese Woldeyohannis
School of Information Science
Addis Ababa University
Addis Ababa, Ethiopia
michael.melese@aau.edu.et

Mesay Gemeda Yigezu
Department of Information Technology
Wachemo University, Ethiopia
Hossana, Ethiopia
messay.gemeda@gmail.com

***Abstract*—In this paper, we described an effort towards the development of parallel corpora for English and Ethiopian Languages, such as Wolaita, Gamo, Gofa, and Dawuro neural machine translation. The corpus is collected from the religious domain and to check the usability of the collected parallel corpora a bi-directional Neural Machine Translation experiments were conducted. The neural machine translation shows good results as a baseline experiment of BLEU score of 13.8 in Wolaita-English and 8.2 English-Wolaita machine translation. The Wolaita-English translation shows a better result than the other pairs of Ethiopian languages and the result of neural machine translation performs well when the amount of dataset increases, thus the amount of dataset has a great impact on the performance. Besides these, the morphological richness of Ethiopian language contributed to the low performance of neural machine translation when the Ethiopian language is used as the target language. Further, we are working on minimizing the effect of morphological richness through different morphological processing techniques in the translation of Ethiopian languages.**

***Index Terms*—Parallel Corpora, Ometo Language, low resourced, Ethiopian languages, machine translation.**

I. Introduction

Language is used by human beings as a means of communication in our day-to-day activities to do various things [1] [2]. These communication includes; giving commands, asking questions and expressing feelings, but we use it specially to communicate information about world. Natural Language Processing (NLP) is a sub field of Artificial Intelligence and getting lot of focus on research and development due to the emergence of its applications in different area [3]. Natural language processing employs computational techniques for the purpose of learning, understanding, and producing human language content [4].

Early computational approaches to language research focused on automating the analysis of the linguistic structure and developing basic technologies such as machine translation, speech recognition, and speech synthesis [4]. A major limitation of NLP today is the fact that most natural language processing resources and systems are available only for technological favored languages, such as English, European and Asian languages. In today's digital world, a lot of written documents are available for technologically favored and resourced languages such as English, European (French, Germany, Italy) and Asian languages (Indian, Chinese, Japanese) [5].

Due to unavailability of NLP resource like machine translation for Ethiopian languages such as Wolaita, Dawuro, Gamo and Gofa, the language speakers are unable to access the resources produced in technologically favored and resourced languages like English. Machine translation is one that helps to benefit the resource deficient language by translating resources from technologically supported languages to the under resourced languages like Wolaita, Dawuro, Gamo and Gofa Wolaita, Dawuro, Gamo and Gofa language. In addition to this, the main aim of translation systems is to produce the best possible translation with minimal intervention hiding the language complexity such as grammar, syntax and semantics of the languages [6].

Ethiopian languages which are under resourced and technologically disadvantaged have limited NLP application due to unavailability of NLP resources [7]. Among these Ethiopian languages, Wolaita, Gamo, Gofa and Dawuro languages which belongs to Omotic language family highly suffer from the lack of language resource to take the advantage of the technological supported language [8]. In addition to this, manual translation is expensive, time consuming, needs professionals, and it is complex to provide the translated material in short period of time [9]. So, there is a need to do automatic machine translation from English to Ethiopian languages to overcome the above stated problems and facilitate the language speakers to access documents written in any Ometo and English language as well as vice versa.

However, machine translation requires parallel or comparable corpora in order to translate from one languages to the other. Therefore, there is a need to collect, pre-process and prepare a language corpus for English and Ethiopian languages for machine translation. Accordingly, this study attempt to collect and prepare a parallel corpus for English to Ethiopian languages for machine translation.

II. Machine translation

Machine translation is one of the initial task taken by the computer scientists and the research in field of NLP for last five decades [10]. For Machine translation, preparation of corpus which contains source and target language dataset are essential to train and test translation models [11], [12], [13]. Several studies and applications have been done for foreign languages using different methodologies and approaches. Most of the machine translation works have been done on language pair of English and other languages, such as Arabic [14], Japanese [15], India [10], Malayalam [16], Bangla [17] are among others. However, research in the area of MT for Ethiopian languages, which are under-resourced as well as economically and technologically disadvantaged, has started very recently [7]. Some of research done for Ethiopian languages are English-Amharic language [5], [18], [19], English-Afaan Oromo machine translation [5], [20], English-Geez [5], English-Wolaita [5] and English-Tigrigna [5], [21]. On the other hand, Gamo, Gofa and Dawuro languages which are predominantly spoken in southern part of Ethiopia are disadvantaged from using available resources on the web due to unavailability of NLP application for these languages. In addition, since there are no standard corpora for conducting replaceable and consistent experiment in machine translation to evaluate the performance.

III. Ethiopian Languages

Ethiopia is a country that has more than 85 languages grouped by linguists into Semitic, Cushitic, Omotic of the Afro-Asiatic and Nilo-Saharan Phyla [22]. The Ometo linguistic group consists of several related languages within the Omotic language family of the Afro-Asiatic phylum [23]. Ometo languages are sub-grouped into North, South, East, and West Ometo [24] The Northern Ometo group includes languages that traditionally have been known as the Wolaita dialect cluster, notably Wolaita, Gamo, Gofa, Dawuro, and Dorze. Wolaita, Gamo, Dawuro and Gofa are spoken in a contiguous territory in the southern nations, nationalities and people's regional state (SNNPR), in an area previously known as North Omo Zone [23]. Recently, North Omo Zone has been further split into four smaller administrative zones, namely, Gamo, Gofa, Wolaita and Dawuro zones.

Wolaita refers to people, language, and the area in the southern part of Ethiopia located in the Wolaita zone with around 2.48 million speakers of the language [25]. Wolayta and Wolaittattuwa are common names for the language. It can also be referred to as Wolaita Doonaa lit [26]. 'moof Wolayta') or Wolaita K'aalaa (lit. 'word of Wolaita') [26]. The Wolaytta language is genetically close to Gamno, Gofa, Dawro, and other languages spoken around the Wolaita zone [27]. The language is given as medium of instruction at primary school level and taught as a subject in secondary and high school. Currently, the language is offered as a program in Bachelor Degree at Wolaita Sodo University. Meanwhile the language is serving as working language and means of communication in government offices in Wolaita Zone.

Dawro is an Omotic language spoken primarily in the Dawro zone of the SNNPR in the Southwest of Ethiopia [28]. Dawro people also refers to their language (locally) as Daurotsua or Dauro K'ala with an approximate speaker of 538,000 [29]. Dawro is mutually intelligible with the neighboring and related languages Wolaita, Gamo, and Gofa. Dawro is, however, the more divergent of the four and the mutual intelligibility is asymmetric in favor of Dawro, meaning that speakers of Dawro have an easier time understanding Wolaita, Gamo, and Gofa [28]. Dawro is used in education in the Dawro Zone, and students receive native language instruction through all grades and now it is also possible to study Dawro in higher education to obtain a diploma [28].

The name Gamo is widely used both as a name of the people and of the language cluster, a collective name to which all the Gamo dialects belong. Gamo language, locally called Gamotstso or Gamotstso doona, is native spoken by more than one million people [30].Gamo is a member of the North Ometo language groups that genealogically descend from Afro Asiatic phylum, Northern Omotic. Gamo is an Ometo language of the Omotic language family used as a language of instruction in the lowest grades in primary school and medium of communication in Gamo Zone and in border areas. Gamo language also spoken in boarder areas of Gamo zone and in Addis Ababa which is capital city of Ethiopia [31].

Gofa is a member of the Omotic language family under the North Ometo Cluster, which is spoken by the people of Gofa as well as the different communities living in the geographical location of Gofa [32]. The people have strong cultural and religious ties with the neighboring people. Some of the languages in contact with Gofa include Oyda, Aari, Dawro, Kontta, Wolaita, and Kucha one of the different varieties of Gamo. These languages have strong influence on the socio-cultural and linguistic identity of the people as well as the development of the language of Gofa in one way or another. Gofa language spoken in Gofa zone and other boarder areas.

A. Writing Systems

The writing system of Wolaitta , Gamo, Gofa, and Dawuro employs the Alphabetic Writing System with an extended version of the Latin Script and consists of 34 letters [22]. Every letter has two forms, one capital and one small, so that the overall number of symbols in the orthography is 68. The four dialects contain certain orthographically important consonants for which the Latin letters do not offer equivalents. The writing system of Wolaita, Gamo, Gofa and Dawuro languages follow Subject-Object-Verb (SOV) word order. They have a strong tendency to use propositions rather than prepositions, to place auxiliary verbs after the action verb, to place a name before a title or to place demonstrative adjectives before the nouns they modify [33].

1) consonants: Wolaitta, Gamo, Gofa and Dawuro share the greatest majority of their consonant inventories [22]. Consonants in the four dialects can be categorized into six

categories: stops, fricatives, affricates, nasals, approximants and semi-vowels. In addition, in all the four variations there is a three-way distinction between voiceless, voiced and ejective consonants for stops, fricatives and affricates in several places of articulation. On the other hand, the phonemic inventories of the four dialects do show certain differences both in terms of number and type [22]. Table I shows the consonant phonemes in Wolaitta, Gamo, Gofa and Dawuro.

TABLE I
CONSONANTS PHONEMES OF WOLAITTA, GAMO, GOFA AND DAWURO LANGUAGES

Language	Consonants
Wolaita	p b p' m w t d n l r D s z s' š c j c' y k g k' ʔh
Dawuro	p b p' m w t d n l r D s z s' š t^s c j c' y k g k' ʔh
Gamo	p b p' m w t d n l r D s z 's š t^s dz c j c' y k g k' ʔh
Gofa	p b p' m w t d n l r D s z s' š t^s c j c' y k g k' ʔh

As shown in Table I Gamo has twenty-six consonant phonemes, Wolaitta has twenty-four consonant phonemes while Dawuro and Gofa have twenty-five phonemes each. Gamo has one peculiar consonant, /dz/, which is absent from the others. Phonemic inventories of the four dialects also show variation with respect to the consonants /t'/ and /s'/. Wolaita has /t'/ while the other three have /s'/ instead.

TABLE II
VARIATION OF WORD FORMULATION IN FOUR LANGUAGES ADOPTED AND MODIFIED FROM [23]

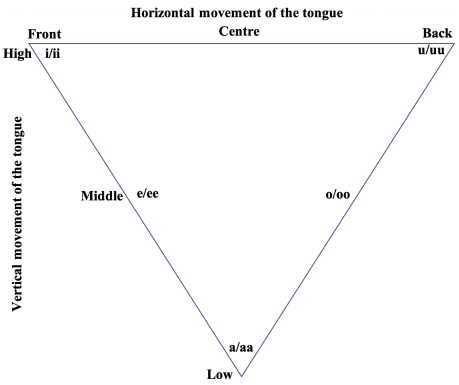
Wolaita	Dawuro	Gamo	Gofa	English
t'ugunta	s'ugunsa	s'ugunsu	s'ugunt^s	'nail'
Sutta	suut^{s}a	Suut^{s}u	suut^s	blood
Heezza	heezza	heedza	heezza	three

Table II shows variation in word formation between four languages. Considering the phonemic inventories of the four dialects, Wolaitta differs much from the other three because, firstly, it lacks the alveolar affricate consonant t^s and secondly, it has its own peculiar phoneme, /t'/ which is absent in other languages [34], [35]. However, the /t'/ in Wolaitta regularly corresponds to /s'/ in the other languages. On the other hand, cognates show that the t^s in Dawuro, Gamo and Gofa corresponds to geminated /tt/ in Wolaitta [34]. On the other hand, Gamo has a unique consonant, /dz/ which is absent in other languages. As illustrated in Table II cognates the /dz/ in Gamo corresponds to z (z) in other languages.

2) Vowels: The vocalic inventory is the same for all the four dialects [23]. Five short and five long phonemic vowels are found in all the four dialects. Like other Latin languages; Wolaita, Gamo, Gofa, and Dawuro languages use five common vowels and Figure 1 shows categories of vowels in the four languages.

Unlike English, Wolaita, Gamo, Dawuro, and Gofa vowels are categorized according to the place of articulation for short and long vowels. These vowels includes the horizontal movement of tongue such as front, center and back and vertical movement of tongue such as high, middle and low as depicted

Fig. 1. Category of Wolaita, Gamo, Dawuro, and Gofa vowels in terms of articulation, adopted and modified from [22]



in Figure 1. The place of articulation of the vowel a/aa is middle and lower, u/uu back and high, i/ii front and high, e/ee front and middle while is o/oo back and middle based on the movement of the tongue in mouth.

3) Numeral system in four Languages: Wolaita, Gamo, Gofa, and Dawuro languages have a quinary numeral system at least historically, the former quinary pattern are widespread across the East and North dialects, which constitute the great majority of the members [23]. At the synchronic level, the quinary system appears to be quite obscured as a decimal one. Table III shows 'one' to 'ten' numeral system for Wolaita, Gamo, Gofa, and Dawuro languages.

TABLE III
NUMERAL SYSTEM IN WOLAITA, GAMO, GOFA, AND DAWURO LANGUAGES ADOPTED AND MODIFIED FROM [23]

	Wolaitta	Dawuro	Gamo	Gofa
1	ʔisso	ʔitta	ʔissio ʔista	ʔistá
2	Naaʔa	Laʔʔa	Namʔa	Namʔʔa
3	Hezza	Hezza	Heedzdza	heedzdzá
4	ʔOyda	ʔOydda	ʔOjdda	ʔOiddá
5	ʔIććaća	ʔIććaća	ʔItʃt ʃatʃtʃa	ʔItʃtʃáʃa
6	ʔUsuppuna	ʔUsupuna	ʔUspuna	ʔUsúppuna
7	Lappuna	Lappuna	Laappuna	Laáppuna
8	Hospuna	Hosppuna	Hospuna	Hósppuna
9	ʔudupuna	ʔudupuna	ʔuddupuna	ʔuddúfuna
10	Tamma	Tamma	Tamma	támma

Numerals 'one' to 'five' are the basic numerals in Wolaita, Gamo, Gofa, and Dawuro. The forms for such numerals are structurally simplex and etymologically opaque. Numerals 'six' to 'nine' exhibit a compounding structure [23]. As shown in Table III, the components in the compound numerals 'six' to 'nine' have undergone drastic phonological changes, and may not be recognized as such at the first sight.

B. Challenges of Ethiopian Languages

Machine Translation is greatly challenged by the linguistic features of the source and target languages. Wolaita, Gamo, Gofa and Dawuro machine translation is challenged by word ordering and morphological complexity of the languages. This languages are morphological rich and follows the same

word orders which is different from English because English follows SVO word order. In addition to this this language uses Compound letters to form words and four languages have also have additional alphabet which does not have corresponding Latin representation , this also challenges the translation when this languages used as target language [22].

IV. Corpus preparation

Compared to a technologically favored language like English, European, and Asian languages resources, Ethiopian languages are under-resourced. Even from the other Ethiopian languages, Ometo languages are difficult to access as most of the data used in these languages are available in printed format.

In this work, different techniques applied to collect, pre-process and prepare parallel corpora for the selected four Ometo languages paired with English. The collected data only fall under the religious domain due to unavailability of language resource to add other domain like legal, historical, health and news. For this study, we collected parallel corpus for Gamo, Gofa and Dawro languages from Ebible[1] which is free website for online Holy Bible that support many languages including Ometo languages. While for Wolaita language we used parallel dataset acquired in the research [35] found in GitHub[2].

To extract the bible data from the website, a web crawler was used for each article after identifying the structure of web documents (HTML) including the page, book, and phrases. Accordingly, Python libraries such as requests, regular expression (RE), and Beautiful Soup (BS) were used to analyze the structure of the websites and extract the content of the article for a given unified resource locator (URL).

After the corpus is collected, the next step is pre-processing collected data, in order to prepare it in a format that is suitable for the different NLP application. Data pre-processing is main step after collecting data that aims to facilitate the training and testing process by appropriately transforming and scaling the entire dataset. To made the dataset suitable for NLP application we applied the following pre-processing steps in collected dataset. All of the collected data was subsequently converted to plain text, clean up from the blank lines and noisy characters, and its encoding was converted to UTF-8 automatically to make it ready to train the system. Converting all of the words into lower case since a word in either uppercase or lower case would be considered as the same word, removing duplicate entries, removing unwanted characters.

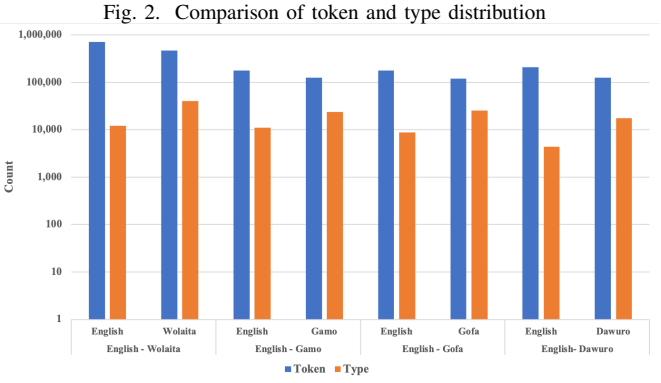
In addition to this, the mapping from one-many, many-one and many-many relationship to one-to-one has been made after all the pre-processing of the document. Table IV presents the detail of collected corpora for Wolaita, Gamo, Gofa and Dawuro languages parallel with English language.

As presented in Table IV, all the Ethiopian languages show an average word length per sentence from 16 to 17 which is far less from English language. Data-set for Gamo, Gofa, and Dawuro are collected from New Testaments of the bible. While the Wolaita dataset is collected from both Old and New Testament of the bible. In addition to this, the size of token and type in English part is much larger than that of the Ethiopian languages. Figure 2 presents the distribution of token and type for English to Ethiopian languages.

TABLE IV
Distribution of sentence, token and type for English-Ethiopian languages pairs.

Language	Sentence	Token	Type	Average words
Wolaita	26,943	469,851	42,049	17
English		703,122	12,131	26
Gamo	7,866	125,509	23,589	16
English		177,410	11,078	23
Gofa	7,928	119,289	25,301	15
English		175,727	8,769	22
Dawuro	7,804	126,734	17,392	16
English		207,954	4,368	27

Fig. 2. Comparison of token and type distribution



As depicted in Figure 2, the change in size of word variants for the Ethiopian language is more than three time except English-Gamo pair which is more than double. The change in token, type and average word per sentence shows the morphological richness of the Ethiopian language over the English. Datasets used for this experiment are freely available[3] along with the source code.

V. Experiments and Discussion

We have conducted bi-directional NMT experiment in collected parallel corpus in order to check the usability of the corpus.

A. Experimental Setup

In order to conduct NMT experiment we divided the dataset in to train, validation and test set. From the total dataset for each language pairs, we used 80/20 train test split and we again split training dataset into training and validation set. We used 70% for training, 10% for validation and the remaining

[1]http://ebible.org

[2]https://github.com/AAUThematic4LT/Parallel-Corpora-for-Ethiopian-Languages

[3]https://github.com/michaelmelese/Ometo-English

20% for testing. To Train and develop machine translation model we used OpenNMT [37] in the Google Colab which allows to write and execute Python in your browser, with zero configuration. Bilingual Evaluation Under Study (BLEU) is used for automatic scoring of the translation result.

B. Experimental Results

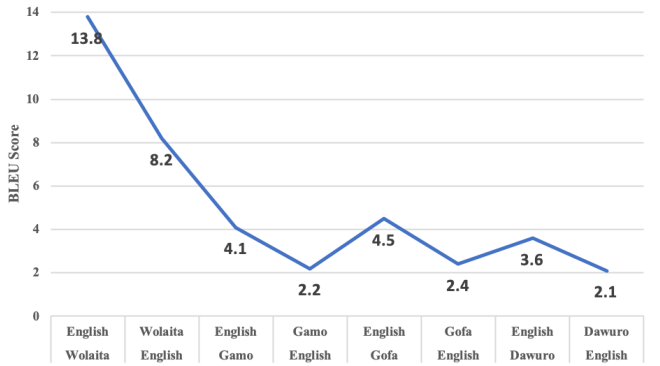
The experiments are conducted in bi-directional one from English to Ethiopian and another from Ethiopian to English language for each Ethiopian language pair. Table V presents the experimental result of bi-directional NMT developed for four Ethiopian languages with English. The results in the

TABLE V
ENGLISH-ETHIOPIAN LANGUAGE NMT EXPERIMENTAL RESULTS

Language pairs	BLEU
Wolaita-English	13.8
English-Wolaita	8.2
Gamo-English	4.1
English-Gamo	2.2
Gofa-English	4.5
English-Gofa	2.4
Dawuro-English	3.6
English-Dawuro	2.1

Table V shows the effect of data size on the performance of NMT systems. As literature supports, the performance of NMT increase as the size of the data increases. Similarly, in the translation of English-Wolaita and Wolaita-English language part, the dataset is four times than the rest of language pairs despite the language difference this is because the dataset used for Wolaita and English language pair contains both old and new testament of the bible data but the other languages contains new testament of the bible. Likewise, the difference in translation performance of three language pairs which have less than English-Wolaita language pair. Figure 3 shows the comparison of BLEU score for bi-directional English to Ethiopian language NMT results.

Fig. 3. Comparison of English-Ethiopian NMT BLEU Score



As depicted in in Figure 3, regardless of dataset size, the model performance is much higher when using English as target language than using English as source language. This is because the language model data favors the English language than that of Ethiopian languages due to morphological richness and complexity of Ethiopian languages. In addition to this, when English is used as source language the translation is challenged by many-to-one alignment.

When we compare the BLUE score of NMT for bidirectional English-Wolaita with other research done for five Ethiopian languages, they used a total of 30,232 sentence from this they used 80% for training, 10% for validation and 10% for testing [5]. The SMT for bidirectional English-Wolaita languages shown 10.49 BLEU Score for English-Wolaita and 17.39 BLEU Score for Wolaita-English. Bidirectional NMT for English-Wolaita shows less result than bidirectional SMT result of English-Wolaita in paper. From this, we can see that NMT model is highly depends on the amount of dataset for a better performance.

VI. CONCLUSION AND RECOMMENDATION

This paper presents the attempt made towards preparing a parallel dataset between English and four low resourced Ethiopian languages spoken in southern part of Ethiopia that belongs to one language family. Parallel dataset is collected from web that contains religion domain and then pre-processed to conduct NMT experiment. Using the collected corpus, a bi-directional neural machine translation experiment has been conducted as a baseline for neural machine translation. The experiment results show that neural machine translation performance depends on the amount of dataset. The morphological complexity is also a factor for NMT performance when Ethiopian languages are used as target language.

To increase the performance of NMT model using large amount of dataset, using different domain with additional linguistic features for Ethiopian languages should be explored in the future.

REFERENCES

[1] A. Al-Hattami, “Strength for Today and Bright Hope for Tomorrow A Phonetic and Phonological Study of the Consonants of English and Arabic,” Lang. India, vol. 10, no. May, pp. 242–365, 2010..
[2] M. Patchell, H. F. Pommer, and W. M. Sale, The Use of Language, vol. 8, no. 1. 1948.
[3] IK. P. Kalyanathaya, D. Akila, and P. Rajesh, “Advances in natural language processing –a survey of current research trends, development tools and industry applications,” Int. J. Recent Technol. Eng., vol. 7, no. 5, pp. 199–201, 2019.
[4] K. Bock and S. M. Garnsey, “Language Processing,” A Companion to Cogn. Sci., pp. 226–234, 2008, doi: 10.1002/9781405164535.ch14.
[5] S. T. Abate et al., “English-Ethiopian Languages Statistical Machine Translation,” Proc. 2019 Work. Widening NLP, pp. 27–30, 2019.
[6] Martinez, L.G., “Human Translation Versus Machine Translation and Full Post-editing of Raw Machine Translation Output.” Dublin City University, Dublin (2003).
[7] S. T. Abate et al., “Parallel Corpora for bi-Directional Statistical Machine Translation for Seven Ethiopian Language Pairs,” October, vol. 0, no. 1, pp. 153–156, 2018.
[8] A. Amha, “The Omotic Language Family,” in The Cambridge Handbook of Linguistic Typology, no. February, 2017, pp. 815–853.
[9] K. Wołk and K. Marasek, “Enhanced Bilingual Evaluation Understudy,” Lect. Notes Inf. Theory, 2014, doi: 10.12720/lnit.2.2.191-197.
[10] S. Saini and V. Sahula, “Neural Machine Translation for English to Hindi,” Proceeding - 2018 4th Int. Conf. Inf. Retr. Knowl. Manag. Diving into Data Sci. CAMP 2018, pp. 25–30, 2018, doi: 10.1109/INFRKM.2018.8464781.

[11] E. Yildiz, A. Cuneyd Tantug, and B. Diri, "The Effect of Parallel Corpus Quality vs Size in English - Toturkish SMT," pp. 21–30, 2014, doi: 10.5121/csit.2014.4710.

[12] P. Zweigenbaum, S. Sharoff, and R. Rapp, "A multilingual dataset for evaluating parallel sentence extraction from comparable corpora," Lr. 2018 - 11th Int. Conf. Lang. Resour. Eval., pp. 3828–3833, 2019.

[13] A. Karakanta, J. Dehdari, and J. van Genabith, "Neural machine translation for low-resource languages without parallel corpora," Mach. Transl., vol. 32, no. 1–2, pp. 167–189, 2018, doi: 10.1007/s10590-017-9203-5.

[14] K. Shaalan, A. Rafea, A. A. Moneim, and H. Baraka, "Machine Translation of English Noun Phrases into Arabic," Int. J. Comput. Process. Lang., vol. 17, no. 02, pp. 121–134, 2004, doi: 10.1142/s021942790400105x.

[15] M. Murata and M. Nagao, "Determination of referential property and number of nouns in Japanese sentences for machine translation into English," pp. 1–8, 1994, [Online]. Available: http://arxiv.orgR. Rajan, R. Sivan, R. Ravindran, and K. P. Soman, "Rule based machine translation from english to Malayalam," ACT 2009 - Int. Conf. Adv. Comput. Control Telecommun. Technol., pp. 439–441, 2009, doi: 10.1109/ACT.2009.113.

[16] R. Rajan, R. Sivan, R. Ravindran, and K. P. Soman, "Rule based machine translation from english to Malayalam," ACT 2009 - Int. Conf. Adv. Comput. Control Telecommun. Technol., pp. 439–441, 2009, doi: 10.1109/ACT.2009.113.

[17] M. A. Al Mumin, M. H. Seddiqui, M. Z. Iqbal, and M. J. Islam, "Neural Machine Translation for Low-resource English-Bangla," J. Comput. Sci., vol. 15, no. 11, pp. 1627–1637, 2019, doi: 10.3844/jcssp.2019.1627.1637.

[18] E. T. Advisor, Y. Assabie, and M. Gebreegziabher, "Bidirectional English-Amharic Machine Translation: An Experiment using Constrained Corpus," no. March, 2013.

[19] M. G. Teshome, L. Besacier, G. Taye, and D. Teferi, "Phoneme-based English-Amharic Statistical Machine Translation," IEEE AFRICON Conf., vol. 2015-Novem, pp. 6–10, 2015, doi: 10.1109/AFRICON.2015.7331921.

[20] S. Adugna and A. Eisele, "English - Oromo Machine translation: An experiment using a statistical approach," 2010.

[21] A. Gebremariam, "Amharic-to-Tigrigna Machine Translation Using Hybrid Approach," 2017.

[22] W. Hirut, "Writing both difference and similarity: towards a more unifying and adequate orthography for the newly written languages of Ethiopia: the case of Wolaitta, Gamo, Gofa and Dawuro," J. Lang. Cult., vol. 5, no. 3, pp. 44–53, 2014, doi: 10.5897/jlc2013.0235.

[23] H. Woldemariam, "Historical notes on numerais in Ometo: The obsolete quinary system," Africa Bibliogr., vol. 2003, no. 1, pp. vii–xviii, 2004, doi: 10.1017/s0266673100007248.

[24] Henok Wondimu, "The Grammaticalization of Copula Markers in the Ometo Subgroup," Addis Ababa University, 2010.

[25] M. Wakasa, "A Sketch Grammar of Wolaytta," 2014.

[26] D. Dalke, "School of Graduate Studies Department of Linguistics Tense , Aspect and Mood (Tam) in Wolayta," 2012.

[27] M. Wakasa, "A Descriptive Study of the Modern Wolaytta Language," The University of Tokyo, 2008.

[28] S. Hanserud, "Dawro verb morphology and syntax," Universtitas Osloensis, 2018.

[29] T. Negese, "Aspect of Dawro Phonology," Addis Ababa University, 2010.

[30] A.Almaz Wasse Gelagay, "The Challenges of Language Standardization: the Case of Gamo," Addis Ababa University, 2018.

[31] A. L. Thomassen, "Contributions to the description of the phonology of the Bonke variety of Gamo," University of Oslo, 2015.

[32] S. C. Hirboro, "Documentation and grammatical description of Gofa," Addis Ababa University, 2015.

[33] M. Wakasa,:A descriptive study of the modern Wolaytta language. Unpublished PhD thesis, University of Tokyo,2008.

[34] W. Hirut, "Notes on the North Ometo dialects:mutual intelligbility tests and structural variations." .

[35] M. L. Bender, Comparative morphology of the Omotic languages, vol. 19. 2000.

[36] S. T. Abate et al., "Parallel Corpora for bi-lingual English-Ethiopian Languages Statistical Machine Translation," Proc. 27th Int. Conf. Comput. Linguist., no. August, pp. 3102–3111, 2018, [Online]. Available: https://www.aclweb.org/anthology/C18-1262.

[37] G. Klein, Y. Kim, Y. Deng, J. Crego, J. Senellart, and A. M. Rush, "OpenNMT: Open-source toolkit for neural machine translation," 20th Annu. Conf. Eur. Assoc. Mach. Transl. EAMT 2017, vol. 1, no. December 2016, p. 22, 2017.