Parallel Corpora Preparation for English-Amharic Machine Translation

yohanens Biadgline, Kamel Smaïli

Parallel Corpora Preparation for English-Amharic Machine Translation*

Yohanens Biadgline[1] and Kamel Smaïli[2]

[1] Bahir Dar Institute of Technology, Bahir Dar, Ethiopia
yohannesb2001@gmail.com
[2] Loria - University of Lorraine, Nancy, France
kamel.smaili@loria.fr

Abstract. In this paper, we describe the development of an English-Amharic parallel corpus and Machine Translation (MT) experiments conducted on it. Two different tests have been achieved. Statistical Machine Translation (SMT) and Neural Machine Translation (NMT) experiments. The performance using the bilingual evaluation understudy metric (BLEU) shows 26.47 and 32.44 respectively for SMT and NMT. The corpus was collected from the Internet using automatic and semi automatic techniques. The harvested corpus concerns domains coming from Religion, Law, and News. Finally, the corpus, we built is composed of 225,304 parallel sentences, it will be shared for free with the community. In our knowledge, this is the biggest parallel corpus so far concerning the Amharic language.

Keywords: Amharic language, Machine Translation · SMT · NMT · Parallel Corpus · BLEU.

1 Introduction

The field of machine translation (MT) is almost as old as the modern digital computer. Throughout these times it undergoes in many technological, algorithmic and methodological milestones [1]. Since its emerging time various approaches have been and being proposed by different researchers in the domain [2,3]. Lexicon (Dictionary) based MT- this strategy for translation depends on entries of a language dictionary. The word's comparable is utilized to build up the deciphered verse. The original of machine translation (late 1940s to 1960s) was altogether in light of machine-readable or electronic lexicons [4–6]. The rule based MT demands various kinds of linguistic resources such as morphological analyzer and synthesizer, syntactic parsers, semantic analyzers and so on [8, 9]. On the other hand, corpus based approaches (as the name implies) require parallel and monolingual corpora [8, 10].

* Supported by Bahir Dar Institute of Technology.

1.1 Motivation

Obtaining accurate translations between two languages using a machine is an open problem. It is expected to be improved in translation accuracy, translation speed, inclusiveness of all languages of the world etc. That is why many researchers and organizations (for example Google, Yandex, Being, Facebook etc.) working hard to create a system that is robust and dependable. But most of the researches ignored the languages that are spoken by people who live in underdeveloped countries. Amharic is one these languages. Even if the performance of MT systems needs improvement; it has been broadly used in the translation sector as an assistant for professional human translators in developed countries. As indicated in [11] it's market share will reach $983.3 million by 2022. So, the main motivation behind this proposal is to contribute our share for the advancement of robust MT system for English-Amharic language pairs and make Amharic one of the language in the market share.

1.2 Machine Translation on English-Amharic Language Pairs

Globally, most of MT researches are done for the languages that are spoken by technologically advanced countries. As a result, a significant improvement has been observed towards development and use of MT systems for these languages. However, MT researches for languages like Amharic (which is considered under resourced) has started very recently. According to literature [8, 12–16] many of English-Amharic MT researches have been conducted using SMT which requires large bilingual corpus. However despite this efforts, still we don't have sufficient amount of digital data to train the SMT model. This shortage of digital data, affects the fluency of the translation model and hence the quality of the translation.

Even-tough Amharic is one of the under resourced languages its counterpart English is the most richest language in terms of data availability. Means, we can find a huge amount digital texts on different resources. In spite of this discrepancies there is a massive need of translation between these languages. News agencies, magazine producers, FM radios, schools, private translators of books, newspaper producers and governmental law announcement paper producers which prints its newspaper in English and Amharic versions are in need of translation on a daily basis. So, we need a MT system to make easy the delivery of courses taught in Ethiopian high schools and universities; to make translation faster and cost effective; to avoid biases in translation; especially, in political domain [14–16].

The major and basic resource required for SMT is a huge parallel corpora [17]. Unfortunately this is not the case for Amharic language. The collection and preparation of parallel corpora for this language is, therefore, an important endeavor to facilitate future MT research and development. We have, therefore, collected and prepared parallel corpora for English-Amharic Languages. This paper will describe an attempt that we have made to collect and prepare English-Amharic parallel corpora and the experiments conducted using the corpora.

1.3 Nature of English-Amharic Language Pairs

Amharic Language Amharic is the second most-spoken Semitic language on the planet, next to Arabic. Of all the languages being spoken in Ethiopia, Amharic is the most widely spoken language. It is the official or working language of the states within the federal system. Moreover, it is used in governmental administration, public media and mass communication (television, radio, literature, entertainment, etc.), and national commerce. Figures change between scientists; notwithstanding, numerous vibe that it has around 57 million speakers. Outside Ethiopia, Amharic is the language of somewhere in the range of 4.7 million emigrants (mainly in Egypt, America, Israel, and Sweden). As of late the number of Amharic talking populace has expanded in Britain and other European nations significantly [18, 20].

Amharic (አማርኛ/əmərɨgnə) is composed with its own script (a variant of the Ge'ez (ግእዝ/*gə'əzzə*) script known as Fidel(ፊደል/*fidəl*) a semi-syllabic framework (Depicted in Table 1). 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. It has 33 primary characters, each representing a consonant and each having 7 varieties in form to demonstrate the vowel which takes after the consonant (Amharic vowels are depicted in Table 2). These 33 sets of 7 shapes are the "common characters"; yet close to them there are additionally various "diphthong characters", each representing a consonant and a following vowel with a */wu/* sound (or, in one case, a */yu/* sound) interposed between them. In composing, none of them is crucial in light of the fact that similar sounds can simply be spoken to by mixes of the customary characters, yet a large number of them are in common use and, in general, they can't be disregarded [14, 20]. Additionally, even if they are not used regularly, Amharic has its own numerals. These are depicted in Table 3 and Table 4.

Both Amharic and the related languages of Ethiopia are written and read from left to right, in contrast to the other Semitic languages like Arabic and Hebrew.

- **Syntactic and morphological nature of the language** Unlike English, Amharic is a morphological complex language. Amharic make use of the root and pattern system [14, 19, 20]. A root (which is called a radical) is a set of consonants which bears the basic meaning of the lexical item whereas a pattern is composed of a set of vowels inserted between the consonants of the root such as in Arabic. Such derivation process makes these languages morphological complex. A derivation process that deals with word-formation; such methods can create new words from existing ones, potentially changing the category of the original word.
  In addition to the morphological information, some syntactic information are also expressed at word level. Furthermore, an orthographic word may attach some syntactic words like prepositions, conjunctions, negation, etc. [21, 22]. In this languages, nominals are inflected for number, gender, etc. At the sentence level Amharic follow Subject-Object-Verb (SOV) word order. On

Table 1: A list of Amharic scripts.

	h	l	h	m	s	r	s	sh	q	b	v	t	ch	h	n	ñ	ä
ɛ, ə	ሀ	ለ	ሐ	መ	ሠ	ረ	ሰ	ሸ	ቀ	በ	ቨ	ተ	ቸ	ኀ	ነ	ኘ	አ
u	ሁ	ሉ	ሑ	ሙ	ሡ	ሩ	ሱ	ሹ	ቁ	ቡ	ቩ	ቱ	ቹ	ኁ	ኑ	ኙ	ኡ
i	ሂ	ሊ	ሒ	ሚ	ሢ	ሪ	ሲ	ሺ	ቂ	ቢ	ቪ	ቲ	ቺ	ኂ	ኒ	ኚ	ኢ
ä	ሃ	ላ	ሓ	ማ	ሣ	ራ	ሳ	ሻ	ቃ	ባ	ቫ	ታ	ቻ	ኃ	ና	ኛ	ኣ
e	ሄ	ሌ	ሔ	ሜ	ሤ	ሬ	ሴ	ሼ	ቄ	ቤ	ቬ	ቴ	ቼ	ኄ	ኔ	ኜ	ኤ
e	ህ	ል	ሕ	ም	ሥ	ር	ስ	ሽ	ቅ	ብ	ቭ	ት	ች	ኅ	ን	ኝ	እ
e	ሆ	ሎ	ሖ	ሞ	ሦ	ሮ	ሶ	ሾ	ቆ	ቦ	ቮ	ቶ	ቾ	ኆ	ኖ	ኞ	ኦ

	k	kh	w	ä	z	zh	y	d	j	g	th	ch	ph	ts	ts	f	p
ɛ, ə	ከ	ኸ	ወ	ዐ	ዘ	ዠ	የ	ደ	ጀ	ገ	ጠ	ጨ	ጰ	ጸ	ፀ	ፈ	ፐ
u	ኩ	ኹ	ዉ	ዑ	ዙ	ዡ	ዩ	ዱ	ጁ	ጉ	ጡ	ጩ	ጱ	ጹ	ፁ	ፉ	ፑ
i	ኪ	ኺ	ዊ	ዒ	ዚ	ዢ	ዪ	ዲ	ጂ	ጊ	ጢ	ጪ	ጲ	ጺ	ፂ	ፊ	ፒ
ä	ካ	ኻ	ዋ	ዓ	ዛ	ዣ	ያ	ዳ	ጃ	ጋ	ጣ	ጫ	ጳ	ጻ	ፃ	ፋ	ፓ
e	ኬ	ኼ	ዌ	ዔ	ዜ	ዤ	ዬ	ዴ	ጄ	ጌ	ጤ	ጬ	ጴ	ጼ	ፄ	ፌ	ፔ
e	ክ	ኽ	ው	ዕ	ዝ	ዥ	ይ	ድ	ጅ	ግ	ጥ	ጭ	ጵ	ጽ	ፅ	ፍ	ፕ
e	ኮ	ኾ	ዎ	ዖ	ዞ	ዦ	዗	ዶ	ጆ	ጎ	ጦ	ጮ	ጶ	ጾ	ፆ	ፎ	ፖ

Table 2: A list of Amharic vowels and their pronunciation.

Vowels	**IPA**	**Translation**	**English Approximation**
አ	*ɛ, ə*	ä (e,eh)	The "e" in set (sometimes a schwa)
ኡ	*u*	u (ou, oo)	The "oo" in foot or soon
ኢ	*i*	i (ii, ee)	The "ea" in seat
ኣ	ä	a (ah)	The "a" in bar
ኤ	*e*	e, é (ie, ié)	Similar to "a" in Way except with no glide
እ	*ɨ*	ə (i, ih)	The "e" in Roses
ኦ	*o*	o (oh or au)	The "oa" in Boat(or the au in maul)

Table 3: A list of Gee'z/Amharic Numbers.

Arabic	1	2	3	4	5	6	7	8	9
Amharic	፩	፪	፫	፬	፭	፮	፯	፰	፱

Table 4: A list of Gee'z/Amharic Numbers.

Arabic	10	20	30	40	50	60	70	80	90	100
Amharic	፲	፳	፴	፵	፶	፷	፸	፹	፺	፻

the contrary, English language uses Subject-Verb-Object (SVO) word-order (Amharic morphology alteration, Amharic syntactic structure and English syntactic structure are depicted by Table 5, Table 6, and Table 7 respectively).

Table 5: An example of Amharic morphology alteration.

Verb	**Derived Words**	**Gloss**	**POS category**
ቀደሰ/*kədəsa*/	ቀዳሲ/ *kədəsi*/	Praise (male)	Adjective
	ቀዳስያን/ *kədəsiyən*/	Praises(they)	Adjective
	ቀዳሲት/ *kədəsit*/	Praise (female)	Adjective
	ቀዳስያት/ *kədəsiyət*/	Praise (they)	Adjective
	ቅዳሴ/ *kidəse*/	Praising/thanks	Noun
	ቅድስና/ *kidisina*/	The act of praising	Noun
	ቅድስት/ *kidisit*/	Praised (female, singular)	Adjective
	ቅዱሳት/ *kidusət*/	Praised (female, plural)	Adjective
	ቅዱስ/ *kidus*/	Praised (male, singular)	Adjective
	ቅዱሳን/ *kidusən*/	Praised (male, plural)	Adjective

Table 6: Amharic syntactic structure

Subject	**Object**	**Verb**
ኢትዮጵያ / *itiyoppya*/	አፍሪካ / *əfirika*/	ውስጥ ናት / *wusitinət*/
Ethiopia	Africa	is in

Table 7: English syntactic structure

Subject	**Verb**	**Object**
Ethiopia	is in	Africa

2 Related works

Different attempts have been made to collect English-Amharic parallel corpus. Below we summarize the researches with most significance to our research. The

most recent attempt to collect English-Amharic parallel corpus is done by Gezmu et al [23]. They have managed to collect 145,364 English-Amharic parallel sentences. The experimental results show that they achieved 20.2 and 26.6 in BLEU score by using Phrase Based Statistical Machine Translation (PBSMT) and NMT models respectively.

Abate et al. [8] collected the English Ethiopian Language (EEL) parallel corpus. They made an attempt to collect parallel corpus for seven major languages of Ethiopia. Amharic was one of them and totally, they collected 40,726 English-Amharic parallel sentences. The SMT approach applied to the collected corpus produced 13.31 BLEU score.

The low resource languages for emergent incidents (LORELEI-Amharic) was developed by the Linguistic Data Consortium and is comprised of a monolingual and parallel Amharic text [24]. It has 60,884 English-Amharic sentences.

As we can observe from the above paragraphs; the largest parallel corpora for English-Amharic language pairs is collected by Gezmu et al [8].

3 Parallel Corpora preparation for the language pairs

A corpus is a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language [25]. Corpus is not any kind of text. It is a sample/collection of texts of a given language which should be representative with regards to the research hypothesis [26]. In this section we will discuss step by step the tasks we have accomplished to collect our bilingual parallel corpora. Work flow of this process is depicted in (Fig. 1).

3.1 Selection of data sources

High quality parallel corpus is crucial for creating SMT or NMT systems [27]. Although high quality parallel corpora is largely available for official languages of the European Union, the United Nations and other organization. It is hard to find enough amount of open parallel corpus for languages like Amharic. So, the only option we have is to create this corpus by ourselves. To do that we should first identify domains with abundant amount of information in a text format. After identifying the domains we collected the raw digital texts from the internet. The collected text data fall under the religious, legal and news domains for which the Amharic text has the corresponding translation in English. Even if there is no shortage of data for English; these are the domains with huge amount of digital text data for Amharic language.

3.2 Collection of Crude Data

In this work, we used different tools and techniques to collect the parallel corpus. As the main tools HTTrack and Heritrix are utilized to crawl and archive

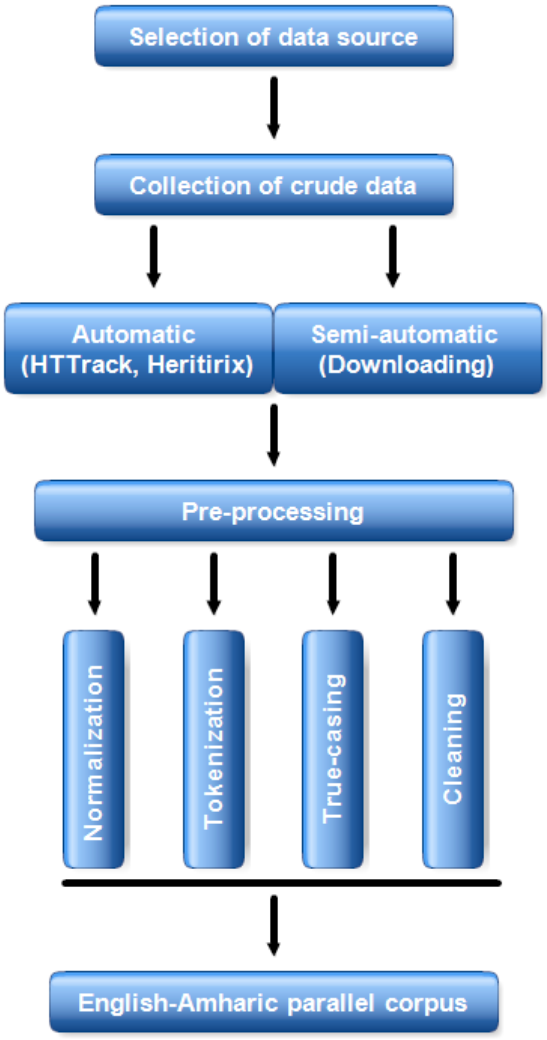

Fig. 1: Process of collecting parallel data.

different websites and news blogs [28, 29]. Additionally, we downloaded a considerable amount of legal documents from different sources. Finally, we extracted the parallelly aligned text data from the collected raw data and merged them into a single UTF-8 file for each language. Still now we have collected a total of 225,304 sentences for each language. Table 8 show detailed information about our corpus.

The total corpus consists of 2,814,888 and 4,068,831 tokens (words) for Amharic and English languages respectively. These figures reveals an interesting

Table 8: Detailed information about the parallel corpus

Domain	**Number of sentences**
Religion	194,023
Law	14,515
News	16,766
Total	**225,304**

fact. In our corpus English uses approximately 1.44 words to express an idea which is written in Amharic with one word.

3.3 Data pre-processing

Data pre-processing is an important and basic step in preparing parallel corpora. Since the collected parallel data have different formats, punctuation marks and other unimportant contents, it is very difficult and time-consuming to prepare usable corpus from it. As part of the pre-processing, unnecessary links, numbers, symbols and foreign texts in each language have been removed. Additionally, character normalization, sentence tokenization, sentence level alignment, true-casing and cleaning are performed [17].

- **Character normalization** there are characters in Amharic that have similar roles and are redundant. To avoid words with same meaning from being taken as different words we replaced these set of characters with similar function into a single most frequently used character. For Example: ሀ, ሐ, ኀ have similar sound and usage in Amharic. They are used to represent sound */hə/*. Similarly, ሠ and ሰ are used to represent the sound */sə/*. So, we removed these characters and substitute them by ሀ and ሰ respectively. Additionally, numerals from Amharic (፩ ,፪ , ፫ ... ፱) to Arabic (1, 2, 3 ...9) have been changed.

- **Tokenization** Tokenization or segmentation is a wide concept that covers simple processes such as separating punctuation from words, or more sophisticated processes such as applying morphological treatments. Separating punctuation and splitting tokens into words or sub-words has proven to be helpful to reduce vocabulary and increase the number of examples of each word, improving the translation quality for certain languages [30]. Tokenization is more challenging when dealing with languages with no separator between words. But this is not the case in this work. Inherently both languages use a word level tokenization. The main task done in this stage was separating words from punctuation marks.

- **True-Casing** we perform this task in order to insure the proper capitalization of every sentence in the corpora. To achieve this we used the Moses

built-in truecaser script. This pre-processing is done only for English Language. Because, grammatically every English sentence should start with an uppercase letter but this is not the case for Amharic. Means, there is not uppercase and lowercase letters in Amharic character map [31].

– **Cleaning** this step is performed to remove empty lines; to avoid redundant space between characters and words; and to cut and discard very long sentences on the parallel corpus simultaneously [32]. At this stage we only consider sentences with 80 words long at most. After performing this task the total number of sentences are reduced to 218,365 sentences from the collected 225,304 Sentences.

4 Experiments and Results

4.1 Experimental Setup

We present two different methods for translation. We used Moses and OpenNMT to train different MT systems. Statistical and neural network based models respectively.

– **SMT experimental setup** creating SMT systems involves two main steps. Creating the language model and training the translation system. A statistical language model is a probability distribution over sequences of words and assigns a probability to every string in the language [33]. Our language model is built with the target language Amharic. We used KenLm to create a 3-gram language model. Totally, 225,304 sentences are utilized for this purpose. After our language model is created the next step was training the translation system. This process enables the model to grasp the relationship between English and Amharic. The model was trained with our pre-processed and cleaned parallel corpus (with 218,365 parallel sentences). As part of the training; word alignment (using GIZA++), phrase extraction and scoring are done. Additionally lexical reordering tables were also created. Then we binarised the model for quick loading at testing stage. Mathematically, the translation model is depicted by equation (1) and (2) where a indicates the Amharic language and e, the English one. Before testing our translation model it should be tuned on other unseen data set. This process enable us to modify the training parameters. These parameters come with default values. However the parameters should be adjusted for each new corpus. In order to tune our translation model we used a distinct small amount of parallel corpus with size of 3121 sentences. This corpus is tokenized and true-cased before it was used.

$$P(a|e) = \frac{P(e|a)P(a)}{P(e)} \tag{1}$$

$$\hat{a} = \underset{a}{argmax} P(a|e) = P(e|a)P(a) \tag{2}$$

Testing is the final stage of our SMT experiment. At this stage we measure how fluent our translation model is. For this purpose we used a distinct corpus of size 2500 sentences. This test set corpus was tokenized and true-cased before it was used. Since our goal is to translate from English to Amharic; we tested our translation model by providing the source language testing corpus (the English sentences). Finally, our translation model translates this English sentences to an Amharic version.

– **NMT experimental setup** for the sake of this experiment we used OpenNMT: Neural Machine Translation Toolkit [35]. The corpus was split as for the SMT experiment into three parts training, validation and testing sets. Then we perform Byte Pair Encoding(BPE). BPE enables NMT model translation on open-vocabulary by encoding rare and unknown words as sequences of sub-word units. This is based on an intuition that various word classes are translatable via smaller units than words. The next step is preprocessing; actually it computes the vocabularies given the most frequent tokens, filters too long sentences, and assigns an index to each token. Training the main and time consuming task in this experiment. To train our NMT model we used Recurrent Neural Networks(RNN) with attention mechanisms. Because, attention mechanism has been shown to produce state-of-the-art results in machine translation and other natural language processing tasks. The attention mechanism takes two sentences, turns them into a matrix where the words of one sentence form the columns, and the words of another sentence form the rows, and then it makes matches, identifying relevant context [36]. This is very useful in machine translation. While we train our RNN model it takes approximately eight and half hours on a GPU equipped device. The detailed parameters of the RNN model are depicted in Table 9.

Table 9: Parameters and values of RNN model

Parameters	**Values**
Hidden units	512
Layers	6
Word vec size	512
Train steps	20000
Batch size	4096
Label smoothing	0.1

4.2 Experimental Results

With this experiment, we created SMT and NMT models for English-Amharic translation. These two languages are different in nature. It means, they are

different in language family, scripts, morphology and syntax. Nonetheless, we build and evaluate our SMT and NMT translation models for the language pairs. We used the BLEU metric to evaluate the performance of our models. The BLEU metric is an algorithm for evaluating the quality of machine translated texts from a source text with reference translations of that text, using weighted averages of the resulting matches. Accordingly, the obtained results are described in Table 10.

Table 10: Comparison of our work with other similar works.

Authors	Year	Sentences	Model used	BLEU
Gezmu et al.	2021	145,364	PBSMT — NMT	20.2 — 26.6
Abate et al.	2018	40,726	SMT	13.31
Strassel et al.	2016	60,884	N/A	N/A
Our work	2021	225,304	SMT — NMT	26.47 — 32.44

Therefore, from Table 10 we can observe that our NMT model shows better translation accuracy than that of the SMT system. The translation accuracy is increased by 22.55%. According to [37] the BLEU score of our NMT model fall between 30 and 40 (actually it is 32.44); which means that the NMT translated texts are understandable to good translations when they are compared with the source texts.

Different attempts have been made to create English-Amharic parallel corpus. Along with that some SMT and NMT experiments are also conducted. For example, Abate et al. [8] collected 40,726 parallel sentences and their SMT model BLEU score was 13.31. Additionally, Ambaye and Yared [38] performed the same (SMT) experiment by using their own corpus and registered 18.74 in a BLEU metrics.

On the other hand even if they are very limited in number, some NMT experiments are also done for the language pairs. Most recently, Gezmu et al. [23] used NMT models and produced 26.6 BLEU score. In [39], the authors collected English-Amharic parallel corpus and conducted NMT experiment on it. As indicated in their research paper the BLEU score ranges between 10 and 12 for different size corpus. Over all by comparing our experiment with the aforementioned attempts, we can say that our research shows an advancement in corpus size and BLEU score for both SMT and NMT.

4.3 Conclusion

MT needs a quite large amount of parallel data. But most of the researches conducted for Amharic language uses small amount of parallel sentences. Their magnitude is measured in terms of thousands and tens of thousands sentences. This is due to the difficulties of finding abundant amount of translated digital

texts from English version. In addition to the size the quality of the available translated documents are not good (can not be directly used for MT purpose). The main objective of this study was to alleviate the aforementioned problem. That is to collect a sizable amount of clean parallel corpus for English-Amharic language pairs. After a prolonged effort, so far we have manged to collect 225,304 parallel and clean sentences. In order to make sure that this parallel corpus is usable for MT or not, we conducted two different experiments. The Results obtained by the two models (SMT and NMT) are promising and our created corpus could be used as a good benchmark corpus which will be proposed for free for the community. Generally, according to the BLEU score interpretation and the results registered by the two models; we can conclude that the prepared parallel corpus is usable for MT researches.

References

1. Slocum, J.: A survey of machine translation: Its history, current status and future prospects. Computational linguistics **11**(1), 1–17 (1985)
2. Antony, P. J.: Machine translation approaches and survey for Indian languages. International Journal of Computational Linguistics and Chinese Language Processing. **18**(1), (March 2013)
3. Hutchins, J: Latest developments in machine translation technology: beginning a new era in MT research. In: Proceedings MT Summit IV.: International cooperation for global communication, pp. 11–34. (1993)
4. Ashraf, Neeha, Manzoor A.: Machine translation techniques and their comparative study. International Journal of Computer Applications **125**(7), 25–31 (2015)
5. Lambert, Patrik, Rafael E., Núria C.: Exploiting lexical information and discriminative alignment training in statistical machine translation. Diss. Ph. D. thesis, Universitat Politecnica de Catalunya. Spain (2008)
6. Poibeau, T: Machine translation. MIT Pres . (2017)
7. Antony, P. J., and K. P. Soman.: Computational morphology and natural language parsing for Indian languages: a literature survey. International Journal of Scientific and Engineering Research **3**, (2012)
8. Abate, Solomon T., et al: Parallel Corpora for bi-Directional Statistical Machine Translation for Seven Ethiopian Language Pairs. In: Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing. (2018)
9. Antony, P. J., and K. P. Soman.: Computational morphology and natural language parsing for Indian languages: a literature survey. International Journal of Scientific and Engineering Research **3**, (2012)
10. Ben Romdhane, A. and Jamoussi, S. and Ben Hamadou, A. and Smaïli, K.: Phrase-Based Language Model in Statistical Machine Translation. International Journal of Computational Linguistics and Applications. Alexander Gelbukh, Dec. (2016) **3**, (2012)
11. `https://www.grandviewresearch.com/press-release/global-machine-translation-market`. Last accessed June 03 2021.
12. Gebreegziabher, M., Besacier, L.: English-Amharic Statistical Machine Translation. (2012)
13. Teshome, E.: Bidirectional English-Amharic machine translation: An experiment using constrained corpus. Master's thesis . Addis Ababa University. (2013)

14. Teferra A., Grover H.: Essentials of Amharic. Rüdiger Köppe, Verlag, Köln. (2007)
15. Daba, J.: Bi-directional English-Afaan oromo machine translation using hybrid approach. Master's thesis . Addis Ababa University. (2013)
16. Saba, A., Sisay F.: Machine Translation for Amharic: Where we are. In: proceedings of LREC, pp. 47–50.(2006)
17. Rauf, S., Holger, S.: Parallel sentence generation from comparable corpora for improved SMT. Machine translation **25**(4), 341–375 (2011)
18. Abiodun, S., Asemahagn A.: Language policy, ideologies, power and the Ethiopian media. Communicatio **41**(1), 71–89. Routledge (2015). https://doi.org/10.1080/02500167.2015.1018288
19. Leslau, W.: Reference Grammar of Amharic. Wiesbaden: Otto Harrassowitz.
20. Yimam, B.: Root Reductions and Extensions in Amharic. Ethiopian Journal of Language and Literature **9**, 56–88. (1999)
21. Gasser, M.: A Dependency Grammar for Amharic. Workshop on Language Resource and Human Language Technologies for Semitic Languages. (2010)
22. Gasser, M.: HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya. Conference on HUman Language Technology for Development. Alexandria, Egypt. (2011)
23. Mekonnen Gezmu, A., NÃ¼rnberger, A., Bayu Bati, T. (2021). Extended Parallel Corpus for Amharic-English Machine Translation. arXiv e-prints, arXiv-2104.
24. Strassel, Stephanie and Jennifer T. 2016. LORELEI Language Packs:Data, Tools, and Resources for Technology Development in Low Resource Languages. In Tenth International Conference on Language Resources and Evaluation, pages 3273-3280.
25. John S.: Corpus Concordance Collection. OUP.
26. Crystal, D.: An Encyclopedic Dictionary of Language and Languages. Oxford: Blackwell. (1992)
27. Dogru, G., Martín-Mor A., Aguilar-Amat, A. : Parallel Corpora Preparation for Machine Translation of Low-Resource Languages:Turkish to English Cardiology Corpora. (2018)
28. HTTrack Website Copier Homepage, `https://www.httrack.com/page/2/`. Last accessed 10 Oct 2020
29. Heritrix Home Page, `http://crawler.archive.org/index.html`. Last accessed 15 Sepetember 2020
30. Palmer, David D.: Tokenisation and sentence segmentation. Handbook of natural language processing. pp.11–35.(2000)
31. Lita, L. V, Ittycheriah, A., Roukos, S., Kambhatla, N.: tRuEcasIng. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 152—159. Sapporo, Japan (2003)
32. Achraf, O., Mohamed J.: Designing High Accuracy Statistical Machine Translation for Sign Language Using Parallel Corpus—Case study English and American Sign Language. Journal of Information Technology Research **12**(2), (2019)
33. Goyal, V., Gurpreet S.: Advances in machine translation systems. Language In India. pp. 138–150 **9**(11) (2009)
34. Daniel J., James H. M..: Speech and Language Processing. Handbook of natural language processing. Draft of October 2, (2019)
35. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: Proceedings of ACL, System Demonstrations. Association for Computational Linguistics. Vancouver, Canada. (2017)
36. Young, T., Hazarika, D., Poria, S., Cambria, E.: Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3), 55-75. (2018).

37. Google Cloud Home Page , https://cloud.google.com/translate/automl/docs/evaluate. Last accessed January 03 2021
38. Ambaye, T., Yared M.: English to Amharic machine translation. The Prague bulletin of mathematical linguistics (2012)
39. Yeabsira A., Rosa T., Surafel L.: ontext Based Machine Translation With Recurrent Neural Network For English-Amharic Translation. In: Proceedings of ICLR 2020. (2020)