Preserving Meher and Woirata Corpus Languages using Neural Machine Translation

Y D Prabowo[1], M Gabriel[2], Nazarudin[3], T G Ratumanan[4], Martinus Maslim[5]

[1]Computer Science Department, Binus Online Learning, Universitas Bina Nusantara, Jakarta, Indonesia

[2]Nuclear Engineering Department, Texas A&M University, Texas, United States of America

[3]Linguistic Department, Universitas Indonesia, Depok, Jawa Barat

[4]Faculty of Teacher Training and Education, Universitas Pattimura, Ambon, Maluku, Indonesia

[5]Informatics Department, Universitas Atma Jaya Yogyakarta, Daerah Istimewa Yogyakarta, Indonesia

E-mail: yulius.denny@binus.ac.id[1], marthen.gabriel@tamu.edu[2], nazarudin.hum@ui.ac.id[3], tanweyratumanan@gmail.com[4], martinus.maslim@uajy.ac.id[5]

Abstract. Research on languages, particularly regional languages, is highly challenging because there is very little or no language corpus available, particularly for Indonesia's regional languages. This project seeks to construct a translation machine for Indonesian in Meher and Woirata languages and vice versa. However, a corpus of Meher and Woirata languages must first be developed to achieve this. The production of this corpus was carried out through field studies. The researcher requested various speakers of this language to translate manually and then compared the results from several translators through focus group talks to identify the appropriate use of words. The outcomes of this translation process are then written in the form of a database of Indonesian-Meher and Indonesian-Woirata language pairings, which will subsequently be utilized as a learning database to create the translation machine. This research collected 714.000 words in the Meher language and 805.000 words in the Woirata language. These results were then employed as a machine translation learning corpus. The translation output carried out by this machine was then validated through direct assessment by speakers of the two languages. The results of this testing indicated an accuracy above 80% for both translation into the Meher language and translation into the Woirata language. Accordingly, it can be concluded that the construction of the Meher and Woirata language corpus, carried out through field research, successfully collected and established a language corpus for these two languages. Apart from that, the experimental results suggested that employing translation algorithms to convert Indonesian into regional languages and vice versa may be carried out and provide translations with acceptable accuracy. This research contributes to establishing the Meher and Woirata language corpus that could be generally accessed.

Keywords: Meher languages; Woirata languages; machine translation

1. Introduction

Presently, the consensus among linguists is that there are approximately 7,000 languages spoken worldwide, with nearly half of these languages potentially facing extinction within the next few generations [1]. Asia has the most extensive language distribution, comprising approximately 33% of the total languages spoken worldwide [2]. About 742 languages are spoken in Indonesia [3]. Indonesians only employ some languages in their daily activities. Additionally, several regional languages are on the verge of extinction due to the absence of indigenous speakers. Today, acquiring regional languages in Indonesia is now of interest to young people. Maluku is one region where a regional language is nearly extinct. Maluku is home to over 130 distinct languages, as stated by [4]. Concurrently, the Maluku Language Office's Mapping team and the Language Development and Guidance Agency of the Ministry of Education and Culture determined the number of active languages. Maluku Province is home to 61 regional languages [5].

Kisar Island, situated in the Maluku Province, is geographically positioned as the farthest island, sharing boundaries with the nations of Timor Leste and Australia [23]. Kisar Island is home to two distinct indigenous communities, specifically Meher and Woirata ethnic groups. These two ethnic groups exhibit notable cultural distinctions, particularly in language. From a linguistic perspective, it is evident that the two ethnic languages, Woirata and Meher, exhibit distinct characteristics [6]. Woirata is classified as a non-Austronesian language, while Meher is affiliated with the Austronesian and Malay Polynesian language families.

Meher and Woirata are two Maluku languages on the verge of extinction. Presently, the number of speakers of these languages does not exceed four hundred individuals, with the majority being elderly. These active speakers, typically around forty years old, represent the remaining Meher and Woirata language users. Consequently, these two languages are classified as endangered traditional languages. Based on the findings of [4], it has been determined that a dialectrometric disparity of around 74% exists between the two languages. The Woirata language is exclusively used in two villages, namely East Woirata and West Woirata villages. In contrast, the Meher language is spoken by most individuals residing on several islands surrounding Kisar Island, such as Letti Island and Luang Island.

Language vitality refers to a language's capacity to discharge its intended communication purposes effectively [7, 24]. As a result, its utilization in the everyday discourse of speakers within the social sphere establishes it as a standard for language preservation [8]. Various factors are considered when assessing the vitality of a language; these include international, national, provincial, educational, developing, threatened, shifting, endangered, nearly extinct passive, and extinct status [9]. Additionally, community aspirations and language vitality prospects are influenced by the sociopolitical contexts in which language varieties are situated [25]. According to [10] research, examining language vitality and its pace of extinction is intricately linked to investigations on language shift, language choice, and bilingualism. Language extinction is a phenomenon that arises when a community of language speakers undergoes a complete transition to a different language, resulting in the abandonment and subsequent disuse of their original language. The phenomenon of language loss exhibits variability among other languages. The capacity to effectively manage external and internal pressures plays a pivotal role in determining the degree of endangerment and potential extinction a language faces.

Numerous scholarly investigations and empirical inquiries have been conducted about the historical development and linguistic concepts associated with regional languages. In Indonesia and other Asian nations, a notable disparity exists between the official language and the vernacular commonly spoken and comprehended in daily interactions. The national language employed in Indonesia, encompassing the island of Kisar, serves as the linguistic medium for the origins of the Meher and Woirata languages. In a broader context, the application of machine learning in the translation of Indonesian to Meher or Woirata has two primary challenges. Research has yet to be conducted on translating Indonesian into the Meher and Woirata

languages. Furthermore, the absence of a translation corpus for Indonesian to Meher and Indonesian to Woirata can be attributed to the limited number of speakers of these languages. The resolution of these two issues constitutes the primary focus of this paper's investigation.

2. Literature Review

The primary focus of this literature review entails two key objectives. Firstly, it aims to investigate the use of neural machine translation in the context of local language translation. Secondly, it seeks to explore strategies for constructing language corpora, particularly in the case of regional languages where data sources are scarce. Despite having a history of at least a century, there needs to be more consensus over the precise definition of corpus linguistics within language structure research. An often-employed characterization of a corpus is "the compilation and analysis of corpora." [11, 12, 26]. However, for this study, researchers adopt the corpus as corpus linguistics concept, which can be succinctly characterized as the scholarly investigation of language grounded in authentic language usage in real-world contexts. As an evidence-gathering technique, corpus linguistics has evolved to improve descriptions of language structures and usage [13, 27]. Like variety, the size of a corpus is also believed, either expressly or implicitly, to play a role in its representativeness [14]. Assessing the magnitude of the link poses challenges. There is a certain degree of correlation between sample size and representativeness. If our corpus were to encompass all instances of a language or its variant, it would inherently possess representativeness. Furthermore, reducing the sample size would not result in an immediate decline to zero regarding representativeness.

In the era of the Internet, the size of corpora is constrained mainly by technical factors. As an illustration, the English language data within the Google N-Grams database is sourced from a corpus of one trillion words, as referenced in the work of [15]. From a quantitative perspective, the figure denotes a substantial magnitude of linguistic input, surpassing the cumulative amount an individual would typically encounter throughout their lifespan. An individual with average reading capabilities may peruse between 200 and 250 words within a minute. Consequently, it would necessitate an uninterrupted reading endeavor spanning between 7500 and 9500 years to consume the entire corpus entirely. Nevertheless, it is essential to note that the corpus represents a minuscule portion of written English. Moreover, linguistic variations are confined to a tiny subset of published written English and need to encompass the contributions of any authentic English speaker.

Determining the necessary size of a linguistic corpus lacks a definitive answer, except for a potentially honest response that is difficult to determine. Nevertheless, two viable solutions exist in practice. A more conservative response is that the sample size should be sufficiently big to encompass a representative collection of examples about the phenomenon being studied, enabling comprehensive analysis. The less modest response entails the corpus's need to be adequately extensive to encompass substantial samples of various grammatical structures, vocabulary items, and other linguistic elements. Considering the growing availability of literature in different languages on the internet, the assertion made in the second answer may not be as boastful as it initially appears.

Two methods exist for searching within a corpus to identify a specific linguistic phenomenon. The first method involves a manual approach, where one reads through the texts within the corpus and records each occurrence of the phenomenon of interest [16, 28, 29]. The second method involves an automated approach, where a computer program executes a query on a machine-readable version of the texts [17, 30]. Numerous initiatives are underway to compile extensive corpora encompassing various web-accessible textual data. The size of these corpora is undeniably remarkable, albeit they generally consist of billions of words rather than trillions.

Nevertheless, the primary justification for their usage lies solely in their magnitude. However, both the individuals responsible for constructing these corpora and those utilizing them must relinquish any notion that they are working with a comprehensive collection of texts. Additionally, they must grapple with the uncertainty surrounding the composition of the corpus, including the specific texts and linguistic

variations it encompasses, as well as the proportion of content generated by English speakers as opposed to automated entities.

In their article titled "Neural Machine Translation of Low-Resource Languages: A Survey" [18], the authors delve into the recent challenges and applications of Neural Machine Translation (NMT) in the context of low-resource languages, specifically indigenous or traditional languages. The article investigates various strategies, including transfer learning, multilingual approaches, and data augmentation techniques, aiming to enhance the translation quality for these languages. Notably, the authors emphasize the importance of addressing the issue of limited data availability in this domain. In a recent investigation by [19], attention was directed toward utilizing monolingual data synthesis to enhance the performance of Neural Machine Translation (NMT) for languages considered minority or local and possess limited parallel corpora. This research introduces a novel approach for improving training data using back translation and language model-based generation. The experimental results demonstrate a substantial improvement in translation quality due to employing this strategy. The present study used this methodology to construct the corpora of the Meher and Woirata languages. Specifically, Indonesian was manually translated into Meher and Woirata languages by a group of proficient native speakers of each respective language.

3. Research Method

In this study, the researcher initially constructed a translation corpus comprising Indonesian to Meher and Indonesian to Woirata. Two primary methodologies exist for creating a bilingual corpus. The first strategy involves automated making a bilingual corpus from online sources, as discussed by [20]. The second approach entails the manual collection of a bilingual corpus, as explored by [21]. This study entailed the compilation of a corpus, wherein individuals from the local community who had native fluency in Meher and Woirata languages were actively involved. The project's linguist provided guidance and supervision throughout the corpus construction process. Once the corpus has been established, it is employed as training data for the translation algorithm that has been constructed. The research conducted in this context substantially contributes to the conservation of the Meher and Woirata languages using artificial intelligence technology. The flow diagram of the entire research process can be seen in Figure 1.

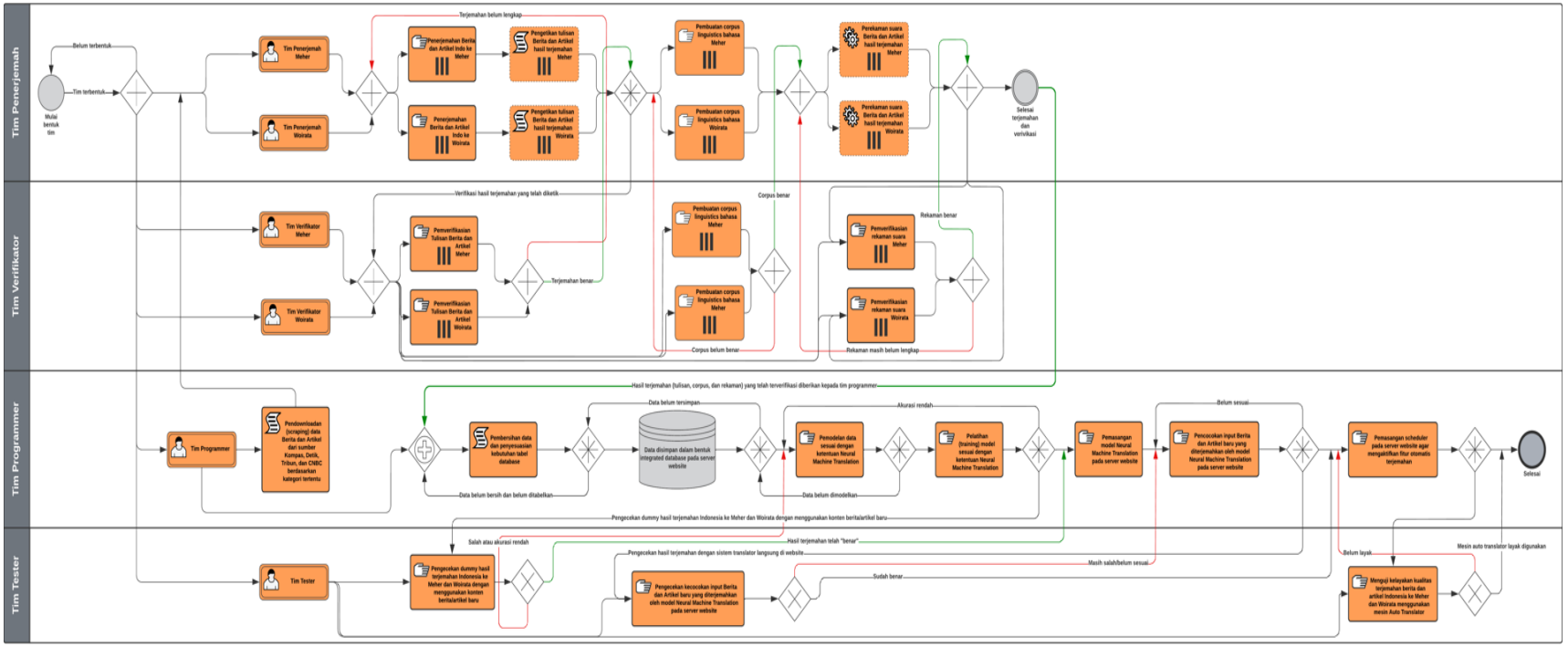

Figure 1. Flow Diagram of Entire Process

The present work involved the construction of a linguistic corpus for the Meher and Woirata languages through the utilization of the crowdsourcing methodology. Crowdsourcing is a phenomenon characterized by seeking contributions or services from numerous individuals to accomplish a particular activity or objective. This process entails decomposing intricate jobs into smaller, more feasible micro-tasks that can be allocated among individual participants. Initially, the study focuses on the project's design, which involves the translation of online news articles written in Indonesian. These articles are retrieved by web crawling and are to be translated into Meher and Woirata languages. The local contributors are then separated into two teams, namely the Meher and Woirata teams, based on their proficiency in the respective languages. The task was subdivided into smaller units, precisely translating news sentences into Meher and Woirata languages. In addition, we have created a quality assurance method that involves doing a post-translation review and engaging three to five participants for each translated item.

In addition to employing the crowdsourcing methodology, we actively engage residents and specialists proficient in Meher and Woirata to obtain an ideal language corpus. The local community is actively involved by promoting translation, writing, and replicating articles in the Meher and Woirata languages, with a particular emphasis on including individuals who have fluency in these local languages. To inspire and incentivize the local populace, we emphasize the cultural value and profound influence stemming from their invaluable efforts in safeguarding this imperiled linguistic heritage. In addition, we have incorporated a feedback mechanism that allows public members to analyze and offer opinions on the collected content. This approach aids us in guaranteeing the accuracy and pertinence of the translation outcomes.

The data utilized in this study comprises articles acquired by web crawling from news websites. Subsequently, this material is transcribed into hardcopy format for the local translation team to disseminate. The translation team comprises Team Meher Translators and Team Woirata Translators. Within each team, some individuals possess native fluency in Meher or Woirata languages. The translation of an article often involves the collaboration of 3-5 translators. In the event of discrepancies or disagreements among the translators, a focus group discussion is conducted to establish a consensus on the preferred translation and the appropriate written and spoken rendition. The translation process outcomes are documented and archived in a tabular structure comprising articles and their corresponding translations. The outcome of this translation is subsequently employed as a replica of the dataset. The corpus collection is later used as training data for the constructed translation algorithm. The translation method employed in this study is founded upon the research completed by [22]. The Neural Machine Translation technique leverages the Deep Decoder architecture and afterward undergoes parameter adjustment to achieve optimal performance based on the available dataset.

4. Results

The present study accomplished the translation of a considerable number of articles in the Meher language and a substantial number of 16,550,000 articles in the Woirata language. Among these articles, it was found that there were 714,000 words in the Meher language and 805,000 words in the Woirata language. This word count is appropriate for capturing the vocabulary employed in everyday resident interactions. Nevertheless, researchers aspire to include extra words and phrases encompassing various contexts, particularly to examine variances in nouns or verbs based on the given context. The given text consists of a single sentence. Figure 2 shows the dataset sample from Bahasa Indonesia to Meher language. The dataset sample from Bahasa Indonesia to the Woirata language can be found in Figure 3.

NEWS TRANSLATION RESULT INDONESIAN TO MEHER				
NEWS IN INDONESIAN			NEWS IN MEHER	
No.	TITLE	NEWS	TITLE	NEWS
1	2	3	4	5
1	Target Pemerintah 2023: Ekonomi Tumbuh 5,9 Persen, Kemiskinan Turun Jadi 7,5 Persen	Liputan6.com, Jakarta Presiden Joko Widodo (Jokowi) pada sidang kabinet paripurna lalu telah menyusun rencana kerja pemerintah (RKP) 2023 untuk proyeksi pertumbuhan ekonomi nasional. RKP tersebut mengambil tema, peningkatan produktivitas untuk transformasi ekonomi yang inklusif dan berkelanjutan. Mengacu pada rencana kerja itu, Menteri PPN/Kepala Bappenas Suharso Monoarfa mengatakan, pemerintah target mencapai pertumbuhan ekonomi 2023 hingga 5,9 persen. "Target sasaran pembangunan dalam RKP tahun 2023 adalah pertumbuhan ekonomi 5,3-5,9 persen," ujar Suharso dalam Musrenbangnas 2022, Kamis (28/4/2022). Tak hanya ekonomi yang tumbuh, pemerintah juga target penurunan tingkat kemiskinan di Indonesia menjadi 7,5 persen. Menurut catatan terakhir Badan Pusat Statistik (BPS), rasio penduduk miskin pada September 2021 sebesar 9,71 persen. Proyeksi lainnya, pemerintah pun memproyeksikan tingkat pengangguran terbuka turun menjadi 5,3-5,6 persen, rasio gini ke level 0,375, penurunan emisi gas rumah kaca 27,02 persen. Kemudian, indeks pembangunan manusia 73,71, nilai tukar petani 103-105, dan nilai tukar nelayan 106-107. Suharso melanjutkan, pemerintah dalam rencana kerjanya juga telah menetapkan beberapa major project yang memiliki peran signifikan dalam mendukung capaian prioritas nasional. "Dalam menyusun major project ini diperkuat dengan penerapan mekanisme clearing house perencanaan untuk menjamin kemanfaatan output pembangunan bagi masyarakat. Sehingga bukan hanya sent, tapi delivered," tuturnya. Hanya saja, kata Margo, jumlah penduduk miskin di tahun 2021 masih lebih tinggi dibandingkan kondisi pada tahun 2019, sebelum adanya pandemi Covid-19. Tercermin dari jumlah penduduk miskin pada September 2019 sebanyak 24,78 juta orang, atau 9,22 persen dari jumlah penduduk Indonesia saat itu. "Kesimpulannya, selama setahun ini penggunaannya cukup baik tapi (angka kemiskinan) lebih tinggi dari kondisi sebelum pandemi," kata dia mengakhiri.	Pemerintah nin target 2023: ekonomi haa 5,9% noro kemiskinan kopur namwali 7,5 persen	Liputan6.com, Jakarta Presiden Joko Widodo (Jokowi) min honopon penepen lalap noro nin manaheli manapu'ik Losno Onno Hair Onno man laa eni horok kanakar honopok nair Losno Hair (RKP riwan woro'o werro'o wokelu roði po'on laa kalari nin nahehe nawo'or popono orrereki Losno Hair. RKP onne nalla mankuku maka (Tema) kikan rana nahehe nawo'or ennen man ma'aruru la nala'a kalari. Nano honorok panaeku honopok nair onnenia, man misla nasala Losno hair nin honorok panaeku na'akeme, Menteri PPN uluakun Bapenas Suharso Monoarfa na'aheni Losno hair nin nahehe nawo'or rala ha'a nahehe nawo'or losno hair riwan ennen riwan woro'o rakan wolima rehi wohii persen. "honorok aki manaonno laa kalari Rahini'i rayapi maa ail RKP raram anna riwan woro'o werro'o wokelu onne nin manha'a wolima rehi wokelu rakan wolima narehi wohii (5,3-5'9) persen ", Suharso na'aheni min Musrembangnas riwan woro'o werro'o woro'o, Kamis (28/4/2022). Ekonomi kan mehe ha'a, Losno Hair me'e nin nahehe nawo'or laa kalari enia roði ra'akopur ðaran lelehe yakyaka/man na'alehe mai Indonesia namwali wo'ikku heruali/7,5 persen, horok man kawali'ur nano Badan Pusat Statistik (BPS) ri morroriana lehelehe yakyaka/mana'alehe lolo September riwan woro'o werro'o ida nin lalap wohii rehi welisuk ida (9'71) persen. Honorok aki man laa kalari namehine Losno Hair po'on laa kalari/ler alam man mai ri morroriana man na'alehe honok naphari kopur namwali wolima rehi wokelu-wolima rehi woneme persen, emisi gas rom koka nin kokopur werro'o wo'ikku rehi woro'o persen. Enamene, hono'ok rahini'i rayapi Ri morroriana nin kilana kuikuil makropropo welisuk wokelu norro welisuk ida, hono'ok man hopok kirna noro man ranoin i'in/nohu mekkieni rahu ida woneme-rahu ida wo'ikku (106-107). Suharso naikar na'awali, losnno hair nin honorok aki manaonno ed'i me'e honopok lalap manasala ulla kikan rakan manamwali losnno hair nin honopok nair honon keren. "Manoði hi'i kakar honopok lalap enia naruri noro hi'i nin kanakar clearing house honorok panaeku leke namwali manoði ra'akene manaku losno hair nina nahini'i nayapi enenn'e laa rimormoriana. Enleke sent mehe ka maa delivered" na'aheni. Maa, Margo na'ahenia, Ri morroriana man na'alehe/lehelehe yakyaka rakan anna riun worro'o welrro'o ida kulu narehi noro ma'ai anna riun worro'o rahu ida wohii, apinha darak (Covid-19) kalla makun po'on nalo lapann'e Ri morroriana man na'alehe/lehelehe-yakyaka lolo wollo September riun worr'o rahu ida wohii law'na makan lalap welrro'o wo'aka, welisuk wo'a ee wohi rehi welrro'o worro'o persen nano Ri morroriana Indonesia do eni e. "Na'akeme manaene namwali anna ennie raram wa'an ma'aruru ma (man lehe yaka/mana'alehe) kulu narehiedi apinha wo'operi (pandemi) kale makun," Ai na'aheni.

Figure 2. Article Translation Sample Dataset from Bahasa Indonesia to Meher

NEWS TRANSLATION RESULT INDONESIAN TO WOIRATA				
Berita Bahasa Indonesia			Berita Bahasa Woirata	
No.	TITLE	NEWS	TITLE	NEWS
1	2	3	4	5
1	Target Pemerintah 2023: Ekonomi Tumbuh 5,9 Persen, Kemiskinan Turun Jadi 7,5 Persen	Liputan6.com, Jakarta Presiden Joko Widodo (Jokowi) pada sidang kabinet paripurna lalu telah menyusun rencana kerja pemerintah (RKP) 2023 untuk proyeksi pertumbuhan ekonomi nasional. RKP tersebut mengambil tema, peningkatan produktivitas untuk transformasi ekonomi yang inklusif dan berkelanjutan. Mengacu pada rencana kerja itu, Menteri PPN/Kepala Bappenas Suharso Monoarfa mengatakan, pemerintah target mencapai pertumbuhan ekonomi 2023 hingga 5,9 persen. "Target sasaran pembangunan dalam RKP tahun 2023 adalah pertumbuhan ekonomi 5,3-5,9 persen," ujar Suharso dalam Musrenbangnas 2022, Kamis (28/4/2022). Tak hanya ekonomi yang tumbuh, pemerintah juga target penurunan tingkat kemiskinan di Indonesia menjadi 7,5 persen. Menurut catatan terakhir Badan Pusat Statistik (BPS), rasio penduduk miskin pada September 2021 sebesar 9,71 persen. Proyeksi lainnya, pemerintah pun memproyeksikan tingkat pengangguran terbuka turun menjadi 5,3-5,6 persen, rasio gini ke level 0,375, penurunan emisi gas rumah kaca 27,02 persen. Kemudian, indeks pembangunan manusia 73,71, nilai tukar petani 103-105, dan nilai tukar nelayan 106-107. Suharso melanjutkan, pemerintah dalam rencana kerjanya juga telah menetapkan beberapa major project yang memiliki peran signifikan dalam mendukung capaian prioritas nasional. "Dalam menyusun major project ini diperkuat dengan penerapan mekanisme clearing house perencanaan untuk menjamin kemanfaatan output pembangunan bagi masyarakat. Sehingga bukan hanya sent, tapi delivered," tuturnya. Hanya saja, kata Margo, jumlah penduduk miskin di tahun 2021 masih lebih tinggi dibandingkan kondisi pada tahun 2019, sebelum adanya pandemi Covid-19. Tercermin dari jumlah penduduk miskin pada September 2019 sebanyak 24,78 juta orang, atau 9,22 persen dari jumlah penduduk Indonesia saat itu. "Kesimpulannya, selama setahun ini penggunaannya cukup baik tapi (angka kemiskinan) lebih tinggi dari kondisi sebelum pandemi," kata dia mengakhiri.	Pemerintah te 2023 ti targete : Ekonomi lausana ye 5,9 Persen ne, Kemiskinan Houte ne 7,5 Persen ne.	Amu : Liputan 6.com, Jakarta Presiden Joko Widodo (Jokowi) tersidang Kabinet paripurna sailana na'a ma-tu Rencana Kerja pemerintah (RKP) ti etere ne 2023 tiye ekonomi lausana ti proyeksi le nasional ti iyar nere. RKP U'nai ye ma-ri anarama'in niye, produktivitas ti pai panhemara le ede ekonomi inklusif ti pai tawa yayani o'o ita ner-nerenhai. Una'a sirwisi kira-kira ti mudwa'a, Mentri PPN/Iyaitapul Bappenas Suharrso Monoarfa enen niye Pemerintah te 2023 ekonomi lausana ti a-kira-kira niye niye 5,9 persen ti mudhemara. "RKP 2023 tawan ti mudwa'a ma-ri a-kira-kira niye sirwisi painana ekonomi ti lausana tiye so 5,3-5,9 persen ti mudhemara. Suharso Musrembangnas 2022 mudwa'a enen ta'i o'one, har kamisi (28/4/2022). Ekonomi yeyen ta so lause he ha, pemerintah tiye ede Indonesia dor-par lapanen ti me'ne ara target ne lapai niye 7,5 persen ti namore. Badan Pusat Statistik (BPS) September 2021 na'a ma-ri enen sailana na'a erasi o'o yawan tour otolosen tiye 9,71 persen. Proyeksi telira ye, pemerintah te ede sirwisi halin lapanen ti proyeksikan to houte ne 5,3-5,6 persen ti namore, kira-kira iyo'onen tiye ma'rane yamoin etun 0,375 iro, ri emisi gas rumah kaca te houte ne 27,02 persen. Al-ten'ni, indeks'e ma-ro sirwisi painana ye 73,71 persen, hala pai-pain ti nemana serlana ye 103-105 persen o'o meti pai-pain ti'i nemana serlana ye 106-107. Suharso al pa'nen niye, Pemerintah te I sirwisi a-kira-kira ti mudwa'a ede major projek iye tarha ti araume, to ethain apte naire nara alan was yayani ne ina'a nasional na'a umar moren ti'i saka'na. "ina'a major projek eteren ti mudwa'a pai rurin tiye clearing house a-kira-kiran ti iyar naire to ina'a yanin output sirwisi pai to momor let tiye lause le ede wata'e. mara-maran te sent yeyeni he ha ede delivered" em'tawane. Tu'urtita, Margo enen niye, 2021 tawan te hihilua asalia otolesen lapanana ti me'ne 2019 tawan onhal covid- 19 asosote nara ono alwas siyatu iyani. September 2019 madomo mudrake nara Hihilua asalia otolesen'e lapanen niye ratu salin 24,78 walwayana ye 9,22 persen. Ethain Indonesia ti hihilua asalia na rekenenara. Ma-ri iwayana tiye, ina'a tawan uwani mudwa'a nairen te ma yani na'ha (otolesen lapanen tiye) onhali pandemi ti na'a mara-maran ti nasoten nara was iyani. Uwe ene'ne I lukun ti uthalama'i.

Figure 3. Article Translation Sample Dataset from Bahasa Indonesia to Woirata

Based on the available data, a model of an Indonesian translator, referred to as "meher" and "Woirata," was developed. The Meher language model was constructed using a training dataset consisting of a maximum of 714 articles, 714,000 words, and a validation dataset. In the process, the Woirata language model was constructed by utilizing a training dataset consisting of 805 words, while the validation dataset had a maximum of 805,000 words. When the model is trained with the same parameter setup, it achieves an accuracy of 86.04% and a loss function value of 14% for the Meher language. Figure 4 shows the training process for the Meher language.

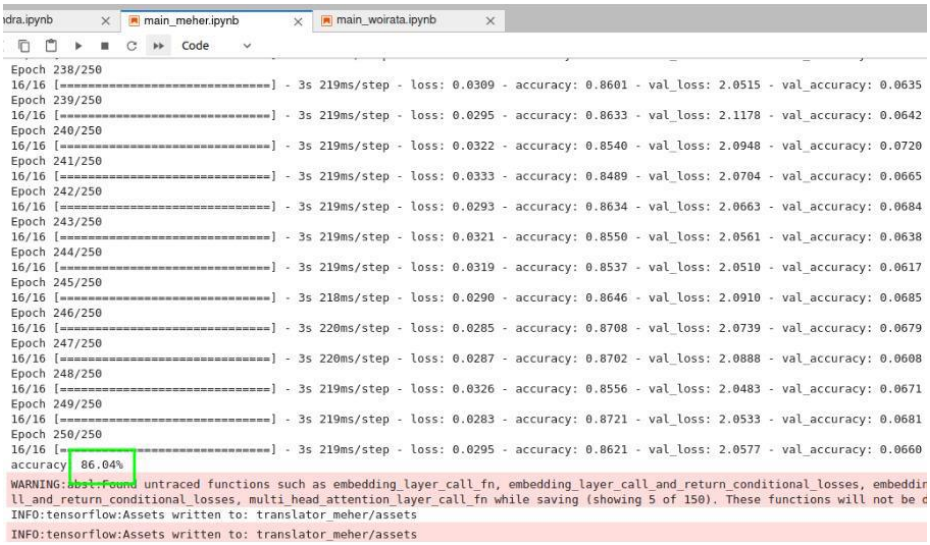

Figure 4. Training Accuracy for Meher Language

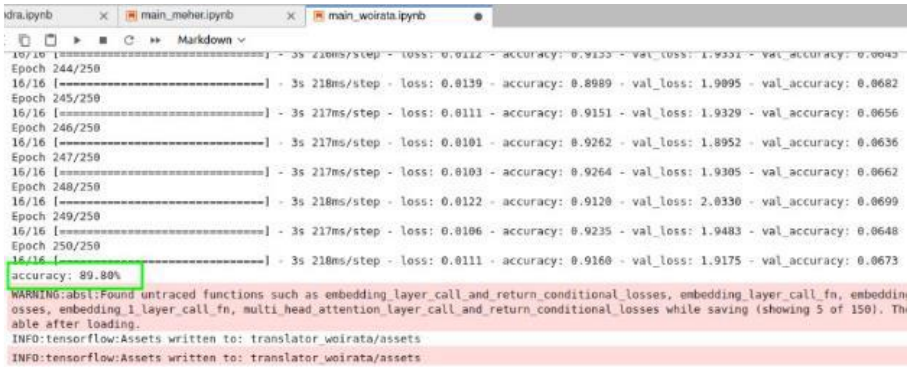

Figure 5. Training Accuracy for Woirata Language

In contrast, for the Woirata language, the model achieves an accuracy of 89.80% and a loss function value of 10%. The result of the training process for Woirata language can be seen in Figure 5. To facilitate users in accurately pronouncing words in Meher or Woirata languages, we supplement them with voice recordings provided by local inhabitants who act as contributors. This practice ensures that the pronunciation of a given term corresponds to its accurate representation in Indonesian. Some examples of voice recordings for Meher and Woirata languages can be seen in Figure 6. The Meher and Woirata languages exhibit variations in pronunciation and intonation that have the potential to alter the semantic content of words. The outcomes of the study were evaluated using a web-based application, which was manually administered to the participants. The decision to do manual testing was made because of the restricted availability of computing devices for accessing website pages.

Figure 6. Example of Recording Pronouncation for Meher and Woirata Language

5. Discussions

There are a lot of challenges that need to be overcome in the subject of language conservation studies for old languages that are in danger of extinction. There is a small number of local speakers, most of whom are older people, and as a result, they require assistance in the technology transfer process. This is the primary difficulty that stands out. The data collection team needs to be comprised of many people who are knowledgeable about technology and can effectively communicate with local communities in their native language. This is because of the problem that has been identified. Developing a conventional language corpus is a challenge that needs to be remedied as soon as possible. Given the limited number of tools that are currently accessible, the building of a linguistic corpus must be done in a relatively manual manner. When it comes to traditional linguistic structures, this is incredibly consistent. The presence of research opportunities in computer science for developing software that can be utilized universally to simplify procedures is brought to light by this challenge—an effort to create a dataset about language.

Traditional language modeling cannot be performed with most currently available language models since they are not algorithmically adequate. Research opportunities for developing language models appropriate for the languages spoken in the region are also made available because of this hurdle. In addition to protecting the language itself, preserving the indigenous knowledge and wisdom ingrained in the cultural traditions of the people who speak the language is also included in the scope of language conservation. There is the potential for a significant contribution to be made by computer technology, particularly the application of artificial intelligence in language. Within the context of Indonesia, however, there are still vital activities that need to be made. It is necessary to implement systematic steps, beginning with constructing a local language corpus, to encourage greater participation of younger scholars concerning this topic.

6. Conclusion

This work has produced a collection of translated writings from the Meher regional language to the Indonesian language from the Woirata regional language to the Indonesian language. This corpus was created by researchers who collected firsthand data, and native speakers translated it. This procedure ensured that the resulting corpus included terminology currently utilized by local language speakers. The construction of two regional language corpora is the most significant contribution that this study has made. These corpora can be used by other researchers doing their studies and can also be utilized in teaching regional languages at the elementary education level.

7. Acknowledgement

The study received financial support from the 2021 Youth Cultural Camp, Cultural Institutions, and Personnel Development under the purview of the Directorate General of Culture, Ministry of Education, Research and Technology. We wish to convey our utmost appreciation and profound acknowledgment on this level.

References

[1] Nazarudin, N., "Causative constructions in Woirata, Kisar Island (Southwest Maluku, Indonesia)," *Wacana, Journal of the Humanities of Indonesia*, vol. 16, no. 1, p. 3, 2016.

[2] Grenoble, Lenore A., "Language ecology and endangerment," in Peter K. Austin and Julia Sallabank (eds), *The Cambridge Handbook of Endangered Languages*, pp. 27-44, Cambridge University Press, 2011.

[3] M. P. Lewis, G. F. Simons, and C. D. Fennig (Eds.), *Ethnologue: Languages of the World*, 17th ed. Dallas, TX: SIL International, 2014.

[4] Summer International Linguistik, *Bahasa-Bahasa di Indonesia*. Jakarta: SIL Internasional Cabang Jakarta, 2005.

[5] Badan Pengembangan dan Pembinaan Bahasa, Kementerian Pendidikan dan Kebudayaan, *Bahasa dan Peta Bahasa di Indonesia*. Jakarta, 2019.

[6] Erniati, E. (2020). "Karakteristik dan Distribusi Fonem Bahasa Woirata di Kabupaten Maluku Barat Daya " [Language Characteristics and Distribution of Woirata Phonems in the District of Maluku Barat Daya]. *TOTOBUANG*, 8(2), 209-223

[7] Candrasari, R., & Nurmaida. (2018). "Model Pengukuran Vitalitas Bahasa: Teori dan Aplikasi pada Penelitian Bahasa-Bahasa Nusantara" (Khalsiah, Ed.). CV Sefa Earth Persada.

[8] Kovanova, E. S. (2019). "Kalmykia and Buryatia: Ethnocultural Security and Language Preservation Issues." *Oriental Studies*, 46(6), 1096–1106.

[9] Lewis, M. P., Simons, G. F., & Fennig, C. D. (2016). *Ethnologue: Languages of the World* (19th ed.). SIL International.

[10] Aritonang, B. (2013). "Vitalitas Bahasa Seget: Kajian ke Arah Pemetaan Vitalitas Bahasa Daerah." *SAWERIGADING*, 19(1), 47-56.

[11] Cheng, Winnie. (2012). *Exploring Corpus Linguistics: Language in Action.* London; New York, NY: Routledge.

[12] Meyer, Charles F. (2002). *English Corpus Linguistics: An Introduction (Studies in English Language).* Cambridge, UK; New York: Cambridge University Press.

[13] Raineri, S., & Debras, C. (2019). "Corpora and Representativeness: Where to Go from Now?". *CogniTextes. Revue de l'Association française de linguistique cognitive*, 19(Volume 19).

[14] Biber, D. (2006). "University Language: A Corpus-Based Study of Spoken and Written Registers." Amsterdam/Philadelphia: John Benjamins.

[15] Brants, T., & Franz, A. (2006). "Web 1T 5-gram version 1." Linguistic Data Consortium. Philadelphia. LDC2006T13.

[16] Crosthwaite, P. (2023). "Corpus Linguistics: Mixed methods research." In C. Chappelle (Ed.), *Wiley Encyclopedia of Applied Linguistics*, 2nd Edition.

[17] Olujimi, P. A., & Ade-Ibijola, A. (2023). "NLP techniques for automating responses to customer queries: a systematic review." *Discover Artificial Intelligence*, 3(1), 20.

[18] Ranathunga, S., Lee, E. S. A., Prifti Skenduli, M., Shekhar, R., Alam, M., & Kaur, R. (2023). "Neural machine translation for low-resource languages: A survey." *ACM Computing Surveys*, 55(11), 1-37.

[19] Wang, X., & Neubig, G. (2019). "Target conditioned sampling: Optimizing data selection for multilingual neural machine translation." *arXiv preprint arXiv:1905.08212.*

[20] Hung-Ngo, Q., & Winiwarter, W. (2012, May). "A visualizing annotation tool for semi-automatically building a bilingual corpus." In *The 5th Workshop on Building and Using Comparable Corpora (Vol. 2, p. 67).*

[21] Ngo, Q. H., & Winiwarter, W. (2012, November). "Building an English-Vietnamese bilingual corpus for machine translation." In *2012 International Conference on Asian Language Processing* (pp. 157-160). IEEE.

[22] Li, Y., bin Abdullah, M. A. R., & Wong, L. Y. (2022, November). "A Systemic Literature Review of Translator's Style in Translation Studies." In *International Conference on Language, Education, and Social Science (ICLESS 2022)* (pp. 80-91). Atlantis Press.

[23] S. Hawkins et al.. (2024). "Earliest known funerary rites in Wallacea after the last glacial maximum." *Scientific Reports*, vol. 14, no. 1, p. 282.

[24] Z. Rohmah and E. W. N. Wijayanti. (2023). "Linguistic landscape of Mojosari: Language policy, language vitality and commodification of language." *Cogent Arts & Humanities*, vol. 10, no. 2, 2023, Art. no. 2275359.

[25] A. Ramponi. (2024). "Language Varieties of Italy: Technology Challenges and Opportunities." *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 19-38.

[26] S. Laviosa and G. Falco. (2024). "Corpora and Translator Education: Past, Present, and Future Check for updates." in *Corpora and Translation Education: Advances and Challenges*, vol. 9.

[27] W. B. McGregor. (2024). "Linguistics: An Introduction," Bloomsbury Publishing.

[28] J. Wu, C. G. Zhao, X. Lu, and T. Jin. (2024) "A rhetorical function and phraseological analysis of commentaries on visuals." *English for Specific Purposes*, vol. 73, pp. 33-45.

[29] N. Tohidi, C. Dadkhah, R. Nouralizadeh Ganji, E. Ghaffari Sadr, and H. Elmi. (2024). "PAMR: Persian Abstract Meaning Representation Corpus." *ACM Transactions on Asian and Low-Resource Language Information Processing*.

[30] S. Zottin, A. De Nardin, E. Colombi, C. Piciarelli, F. Pavan, and G. L. Foresti. (2024). "U-DIADS-Bib: A full and few-shot pixel-precise dataset for document layout analysis of ancient manuscripts." *Neural Computing and Applications*, pp. 1-13.