Participatory Translations of Oshiwambo: Towards Culture Preservation with Language Technology

Wilhelmina Onyothi Nekoto[*], **Julia Kreutzer**[*a], **Jenalea Rajab**[*c],
Millicent Ochieng[*d], **Jade Abbott**[*e]
[*]Masakhane NLP, [a]Google Research, [c]University of Witwatersrand,
[d]Microsoft Africa Research Institute, [e]Retro Rabbit

Abstract

In this paper, we describe a participatory, collaborative, and cost-effective process for creating translations in Oshiwambo, the most widely African language spoken in Namibia. We aim to (1) build a resource for language technology development, (2) bridge generational gaps in cultural and language knowledge, and at the same time (3) provide socio-economic opportunities through language preservation. The created data spans diverse topics of cultural importance, and comprises over 5,000 sentences written in the Oshindonga dialect and translated to English, the largest parallel corpus for Oshiwambo to-date. We show that it is very effective for machine translation, especially when combined with transfer learning.

1 Introduction

To determine the endangerment of a language, UNESCO (2003) proposes several factors, one of them is *intergenerational language transmission*, describing how a language is passed from one generation to the other. The focus language of this paper, Oshiwambo, has over 1M native speakers and is not classified as endangered overall. Nevertheless, the intergenerational transmission is considered “unsafe” on the UNESCO scale, threatening the future of the language and the culture it is embedded in.

Language technology can help to document and revitalize languages (Bird, 2009; Cruz & Waring, 2019; van Esch et al., 2019; Neubig et al., 2020), but due to its data-centricity, research has to start at the data creation, before considering training and deploying tools that could help to increase the use of the language. In this paper, we propose a *participatory, collaborative, and cost-effective process for creating data that is grounded in and representative of the Owambo culture.* Beyond the immediate purpose of creating an Oshiwambo resource for Natural Language Processing (NLP) research, neural machine translation (NMT) in particular, the aim of our participatory data creation project with Oshiwambo speakers is to *bridge generational gaps in cultural and language knowledge* that have grown through alienation, urbanization and segregation (Section 2). From a socio-economic perspective, we aim to create community and youth opportunities to increase household income, especially for households most affected by the COVID pandemic.

Concretely, we describe a paid data creation workshop with Oshiwambo speakers (Section 3), relating design decisions to the history and the present of the Oshiwambo language and culture, which we review in Section 2. The created data set spans diverse topics of cultural importance, and comprises over 5,000 sentences written in the dialect Oshindonga with English translations, and is the largest parallel corpus for Oshiwambo to-date (Section 2.2). We furthermore show how effective the high-quality data is for NMT, especially when combined with transfer learning (Section 4).

We hope that this initial project can be a pilot for future participatory data creation initiatives, to leave a lasting benefit across generations of the involved communities, where income opportunities go hand in hand with culture and language preservation (Section 5).

2 OSHIWAMBO

Oshiwambo is a Bantu language spoken by the Owambo people, mostly in the North of Namibia and the South of Angola. There are eight dialects of Oshiwambo, only two of them are standardized written dialects, Oshindonga (`ng`/`ndo`) and Oshikwanyama (`kj`/`kua`) . Oshiwambo is only recognised as one of thirteen national languages, and is not the official language of Namibia (which is English), despite being the most widely spoken native language (50% of the population which is 2.5M in 2022)[1]

Without significant effort, the Oshiwambo language, and the Owambo culture it carries, will soon be endangered (UNESCO, 2003), despite the large number of native speakers. In Section 2.1 we discuss how the language's development and transmission has been inhibited throughout the last century (Fredericks, 2007), and how the knowledge holders today are disconnected from younger generations, who predominantly speak English and other languages. Section 2.2 follows with a review of the digitized resources that exist of Oshiwambo today.

2.1 LANGUAGE DEVELOPMENT ACROSS GENERATIONS

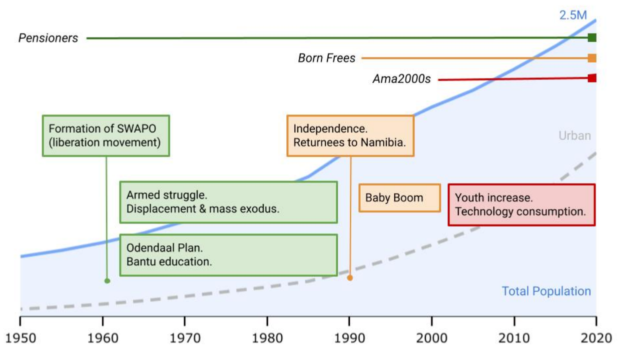


Figure 1: Timeline of Namibia's recent history with selected events, generations and the development of the population size. The demographic statistics are sourced from Worldometers.

Figure 1 highlights the most relevant events in the past 70 years of Namibia's history that has shaped the three current generations of Namibians today. Our summary of Oshiwambo history below, is informed by discussion with elders and veterans (Section 3.1), to reflect their realities and experiences with their language.

The Pensioners are the main knowledge holders today, after having gone through decades of segregation, war, exile and struggle. The South-African regime that followed Namibia's colonization by Germany (1894–1915) implemented the Odendaal Plan in the early 1960s, separating ethnic groups and assigning them separate districts ("homelands"), under the pretext of avoiding ethnic rivalry for the development of the country, but in reality as a measure of control and manipulation (United Nations Department of Political Affairs & Decolonization, 1977; Botha, 2022).

The segregation introduced a hierarchy, where the Owambo's ranked last and were discriminated against, seen as uncivilised. Furthermore, the education system was segregated: Black (Bantu) and White education differed, and the career opportunities for Bantu were limited (commonly nursing, teaching and gardening, while the medicine and science curriculum, e.g. Linear Algebra, were not part of the Bantu education) (Gallo, 2020). Primary schools were taught in Bantu languages specific to the ethnic district, secondary schools in Afrikaans or English (Stell, 2021), which eventually led to youth protests in 1976.

To end this discrimination and segregation under South-Africa's apartheid rule, many young and able-bodied Namibians (30–40% of the population, with a majority from Northern Owambo and Kavango Namibia) joined the liberation movement (Owambo Peoples Organisation, OPO, later renamed to South West Africa People's Organisation, SWAPO) to fight for independence over 3 decades. Many others left the country and lived in exile for education, and hope for a better life. While over 8000 known soldiers perished in the war (SWAPO., 1996), the rest returned with Namibia's independence in 1990, when English was chosen as the only official language (Frydman, 2011).

[1] `https://www.gov.na/languages-spoken`, all URLs accessed on 16 Feb 2022.

The development of Oshiwambo in the country was not taking place: The older generation spent their prime years fighting in the armed struggle, instead of developing the language further, e.g. capturing the development of new terms for new technology that came with industrialisation, or the earlier history in writing.

In exile, other languages, mainly English, had to be learned and were dominating life and education. Upon returning to Namibia, the majority settled in the fast-growing and now multi-ethnic cities, away from the cultural roots of their native language and villages (Mwase, 1990). The importance of their mother-tongue diminished in their everyday life, as English and Afrikaans were dominating in work, administration, and communication , contributing to Oshiwambo continuing to be seen as inferior or archaic. Traveling to the homelands, where the younger generation can immerse in culture, family, and reconnect to their roots, remains expensive for many.

Independence was followed by a baby boom (the Born Frees). A significant increase in technology consumption shaped the newest generation, the "Ama2000s", born after 2000.[2] These generations were born into a completely different Namibia (Smit, 2012). For finding employment, English became the language of the work place, as well as the language of technologies and information. Currently Oshiwambo is taught in some (mostly rural) schools in the first three years of school (as are other native languages), but afterwards the language of instruction is English (Ashton et al., 2008). Oshiwambo spoken at home is not learned sufficiently, especially in a home with full-time employment, leaving a few hours for family time. As such the language and connection to culture get diluted.

The generation of knowledge holders is ageing and at high risk of the devastating effects of the COVID pandemic and other diseases. They have very low retirement income or repatriation benefits, and very few initiatives and platforms exist for them to share and document their experiences and knowledge (Mwase, 1990). Therefore, we engage them in our preliminary studies to share their stories, and as experts in our data creation process. The younger participants of the data creation project also learn from their perspectives and knowledge, and develop re-connections with their native language and culture.

2.2 Existing Digitized Data

For NLP research, Oshiwambo can be considered *extremely low-resourced* and *understudied*: Joshi et al. (2020) classify Oshidonga as "Left-Behind" based on the resources found. Three morphology/tonology-focused studies (Yli-Jyrä, 2011; 2013; 2015) can be found, based on the formalization of the tonal grammar of Oshikwanyama (Halme, 2004).

OPUS lists only one parallel corpus for Oshiwambo (Tiedemann, 2012), the multilingual QED corpus (Abdelali et al., 2014), which claims to contain translations into Oshikwanyama, but they are in fact German. The Universal Declaration of Human Rights (UDHR) corpus (Vatanen et al., 2010) contains Oshiwambo written in Oshindonga, but it spans only 70 sentences.

The Bible has been translated to Oshindonga and Oshikwanyama in 1891–1974 by German and Finnish missionaries and indigenous translators (Ngodji, 2004), but the quality of the translation has been criticized.[3] A new Bible translation project started in 2017 and was estimated to take 12 years and N$25 million (around 1.7 million USD) with a total of seven translators. The cost of the translation of one of the approximately 31k Bible verses is N$300 (around 20 USD).[4] This gives a sense of the effort required for manual translation of existing texts.

Morgan et al. (1991) created a dictionary for Namibian Sign Language (NSL) and English and Oshiwambo of around 580 signs. What this project has in common with ours, beyond the languages, is that the data set creation is rooted in the everyday life and culture of the native speakers, and discussions among speakers are used to develop a consensus on language use.

In 2017, the Namibian constitution was translated from English to Oshiwambo (Oshikwanyama dialect) and is freely available, so we parse, semi-automatically align and prepare it as a parallel cor-

[2] https://mediaupdate.co.za/media/148819/ama-2000--who-are-they

[3] https://economist.com.na/22037/headlines/new-oshiwambo-bible-translation-will-take-12-years-and-n25-million-to-complete/

[4] https://nambible.org.na/translation/

pus (details in Appendix A.1).[5] It contains around 800 sentences, which we use as auxiliary training data (Section 4). A concurrent work to ours (Adewumi, 2022) translated 1k English questions and answers around touristic bookings into Oshindonga. The authors kindly allowed us to use these as auxiliary training data, we will refer to it as *Q&A* dataset. Similar to the constitution data, we expect a mismatch in domains, since the data was not sourced in a context of primary cultural relevance for Oshiwambo speakers.

Oshiwambo data on the web is sparse, because it is an oral vernacular, and because of the generational, cultural and geographical gap between language knowledge holders and digital natives. There are at least four radio stations that purely broadcast in Oshiwambo, and the biggest Namibian newspaper has an Oshiwambo section.[6] These Namibian-centric digital resources could eventually be leveraged for developing language identification models (Caswell et al., 2020), pretraining multilingual models such as AfriBERTa (Ogueji et al., 2021), and training speech recognition models (Doumbouya et al., 2021; Carmantini et al., 2019) as stepping stones to other NLP and speech processing applications.

Despite the scarcity of currently available parallel data for training NMT models for Oshiwambo, significant translation efforts have been invested, e.g. for translations of educational material for Namibian schools (Shatepa & Shikesho, 2019) or Oshiwambo proverbs (Hasheela, 1993), but the resources are not freely accessible in an easy-to-parse format, a fate shared with many other low-resourced languages ($\forall$ et al., 2020). In general, this does not mean that "the language does not have any resources" — they just cannot be found in digital NLP data catalogues, but are rather held by language knowledge holders (Bird, 2020; Hämäläinen, 2021). This motivates our approach of resource creation from a *participatory* angle and with a focus on intergenerational transmission (Winschiers-Theophilus et al., 2010; Bird, 2020; $\forall$ et al., 2020).

3 The Data Creation Process

The goals for this data creation project were threefold:

1. From the *data growth* perspective, the goal is to yield as many adequate and representative sentences as possible in a short amount of time.
2. From the *cultural* perspective, the goal is also for the participants to think and discuss socio-economic issues in Oshiwambo. This includes, e.g., understanding the role of women in the Oshiwambo tradition, re-imagining drought relief and food parcels from a cultural perspective in rural and informal settlements, Namibian history, the role of endangered species in the traditional beliefs and outwardly to the environment, breaking stigma around HIV and Aids, or traditional home remedies.
3. For *personal development* of the participants, we want to enhance their skills in terms of creative thinking, analytical and technical abilities such as typing or conducting online research of facts, team work and communication.

The pursuit of these goals is reflected in the workshop design: Brainstorm sessions and collaborative discussions served the cultural understanding, team and communication skills; a competition over short sprints for sentence writing made the data generation efficient; and the overall process was very flexible to adapt to the participants present on a given day and their interests and needs.

3.1 Preliminary Work & Feasibility Studies

The design of the workshop was developed over months of exploration, discussions with knowledge holders and feasibility studies, and is part of an ongoing larger-scale mission around language and culture conservancy in rural areas, expanded on in Section 5. Previous explorations and small-scale projects evolved around the following:

1. *Namibian Names*: The project aims to identify, re-identify and/or define the Namibian by the meaning of their traditional names, as a representation of their culture, lineage and family, in

[5] https://www.kas.de/en/web/namibia/veranstaltungsberichte/detail/-/content/kas-uebersetzt-namibische-verfassung-auf-oshiwambo

[6] https://www.namibian.com.na/The-Namibian/Oshiwambo

order to encourage cultural appreciation. Given names in Oshiwambo are crowd-sourced, then their meaning is translated into English, and further sentences are constructed around it.

2. *Human-Environment Relationships*: The strong communion between human and environment is important to the living indigenous groups that rely solely on the land. This project crowd-sources and translates Namibian fauna and flora names into Oshiwambo and Khoi in collaboration with rangers and Hunter Gatherers. It further captures the associated beliefs, and discusses totems as a representation of identity and lineage.
3. *Veteran Stories*: Oshiwambo interviews of veterans centered around combat names, the exodus into exile and life in exile are being recorded, transcribed, and translated.

3.2 Workshop Setup

Participants Requirements for participation were to be either students or unemployed, and equipped with a great command of the Oshindonga dialect and English from high-school, and computer literacy. We focused on recruiting young job seekers in order to provide an opportunity of income and to inspire them. The recruitment via an open call attracted a total of eleven participants (six female / five male) of ages 22 to 41 with expertise in nursing, law, teaching town planning and mechanics. The demographics of participants influence the data generation process, as they are knowledgeable in the language, but would construct simpler sentences than e.g. veterans telling their life stories.

Environment Drawing inspiration from the Deep Learning Indaba[7], we knew that it was key to create an atmosphere that was tranquil and allowed for creative thinking and collaboration, to produce high data quality. Many participants live in informal areas, with limited connectivity, energy and no access to a conducive working space. We chose a venue which would encourage outdoor meditation activities in between breaks, and an indoor conference-style area with a white-board and projector. Nourishment and safety was also important. Catering was provided for participants, which included an all day coffee station, water, soft drinks, tea breaks and lunch.

Costs The total cost of the participatory workshop amounted to N$40k (around 2.6k USD). Participant remuneration was a daily fixed rate based on the participants occupation/education and the market-related minimum wage, but the time requirement was flexible. The provision of the venue and lodging (31%), transport (6%), and food (35%) was considered indirect remuneration, and incurred the largest costs. Stationary (1%) and mobile broadband data (2.5%) took smaller shares. Participants worked on their own laptops or on the organizer's one, so that no new technical equipment had to be purchased. 80% of all meals were sourced from small home-run businesses to support Black-owned businesses, mostly run by young self-employed women. The overall costs are low in comparison e.g. with the Bible translation project (see Section 2.2). However, preliminary work and feasibility studies incurred additional costs, that we did not list here.

3.3 Schedule

The brainstorming and data creation sessions were conducted for eight days in a row, from 9AM to 7PM, with one day of break. Four to eight participants were involved every day. They received an initial training in the evening before the start of the event, where they were given translations of 400 Oshiwambo names (see Project (1) in Section 3.1), and had to create 15–20 parallel sentences using any of the names provided. Participants were also provided with 16 audios from Ololo Nam[8] videos which they could transcribe and translate.

The first day was largely spent on on-boarding, and the following days were partially in a classroom setup, partially virtual. The supervision by the project lead was necessary in the beginning to teach and develop the creative process, but from the second day the participants also worked in a self-supervised manner (with the lead reachable by phone if needed). The participants appointed a leader among themselves to keep them on track and moderate. This was done to ensure that they would have enough room to think creatively and to develop ownership of the process.

[7] https://deeplearningindaba.com/

[8] https://www.youtube.com/channel/UCXKY_Etc8NRzIsxaCgHW_WA

Dataset	Translation Direction	Year	Size #sents (# src tokens / # trg tokens)	Type-Token Ratio src / trg	Avg Length src / trg	Length Ratio	% Singletons src / trg
WON	ng→en	2022	5419 (24.1k / 31.0k)	0.33 / 0.17	4.5 / 5.7	1.34	65.4 / 55.8
Constitution	en→kj	2017	783 (21.1k / 18.6k)	0.14 / 0.24	27.0 / 23.9	0.95	51.9 / 60.8
Q&A	en→ng	2021	1000 (13.0k / 9.0k)	0.16 / 0.31	13.0 / 9.0	0.71	53.4 / 68.1

Table 1: Characteristics of the translations between Oshiwambo and English. The second column indicates the translation direction (src→trg).

By varying the themes, topics and modalities throughout the week, participants remained interested and the coverage of a diverse set of culturally relevant topics was ensured. On day five, for instance, veterans were invited to provide answers to the participants' questions around traditional processes. Day six was focused on extracting verbs and pronouns from local news as seed words for subsequent sentence creation. And on day seven, brainstorming was centered around totems (see Project (2) in Section 3.1), participants introspection, and around the effects of economic systems on the Namibian society, while the last day was spent on free writing and profession-based Q&A. Table 3 in Appendix A.2 describes the setup and activities for each day in detail. Overall, the sentence creation process was slower than expected, and interventions such as the introduction of competitions were introduced to increase the productivity.

3.4 Sentence Creation Process

The data creation is performed in sprints of one hour. Each of the sprints has one "slow", collaborative introduction phase, and then fast intervals of competitive sentence creation with short breaks.

In the first *familiarization and discussion* phase, participants discuss and debate to form a consensus on the writing and the use of keywords around a given topic. These keywords were often seed words proposed by the workshop leader, and then expanded during the discussion and brainstorming. On some days they were obtained from external sources, such as radio transmissions (see Figure 2 in Appendix A.3). If needed, professional language teachers or elders were contacted by telephone to resolve disagreements around certain words. The discussion helps everyone to align their ideas, and clarify uncertainties that could lead to incorrect use or slow sentence creation.

In the next phase, the participants compete for *sentence creation*. Each participant works on one word each for two minutes, writing down as many sentences as possible. The sentences should contain the seed word, and reflect a natural context and usage for this word. The topic can be chosen freely, but the participants are instructed to create complete sentences. On average, they create three sentences per interval. They either directly write down the English translation with the Oshindonga source, or they complete them in the break between intervals. After each interval, each participant reports the number of sentences they completed. Additional incentives were introduced to promote the competitive spirit, such as prizes for the top two participants with the cumulative highest score.

After each day, the project leader revises all collected sentences to improve instructions for the next day. Finally, the entire collection is reviewed to ensure all sentences are in the Oshindonga written, and not spoken dialect. Appendix A.5 provides further reflections on the workshop design.

4 Results

4.1 Data Statistics

Our new data is from here on referred to as WON (Writing Our Narratives). The WON data comprises 5419 relative short sentences. The data creation workshop yielded 13k sentences in total, but the remaining portion is still being revised. Table 1 presents the characteristics of the parallel data quantitatively. The English side is composed of slightly more tokens since it is morphologically poorer than Oshiwambo. The 783 translated sentences of the constitution are around 5× longer, because they contain more formal language. The Q&A data (Adewumi, 2022) adds 1000 sentences of medium length, representing short conversational interactions.

The Type-Token Ratio (TTR) is overall low, and the singleton rate high, meaning that the data sets have many infrequent tokens. This is expected, since diverse topics are covered in relatively few

Source Language(s)	**Training Data**	**BLEU**
Training from Scratch		
`ng`	WON+	8.92
Zero-Shot Testing		
`kj`	Constitution	0.02
`sw`	JW300	0.22
`zu`	JW300	0.26
M2M	Web crawls	0.35
Bantu Transfer and Fine-Tuning		
`kj`→`ng`	Constitution→WON+	5.97
`sw`→`ng`	JW300→WON+	17.64
`zu`→`ng`	JW300→WON+	14.80
Massively Multilingual Transfer		
M2M→`ng`	Web crawls→WON+	24.21

Table 2: BLEU scores (Post, 2018) for models tested on `ng-en` WON translations. WON+: WON data combined with Q&A data.

sentences. Furthermore, there is a marginal vocabulary overlap between data sets despite dialect and domain differences: 4.6% of the Oshindonga word types in the WON data also occur in the Oshikwamyana constitution data, and 9.5% of the English words types. For the Q&A data (Adewumi, 2022), that is written in the same dialect, but covers foreign-sourced data, the overlap is slightly higher: 8.9% of WON Oshindonga word types, and 11.0% of English word types are shared.

4.2 Effectiveness for NMT

Building adequate NMT models in the low-resourced data setting is a continual challenge for many African languages (Martinus & Abbott, 2019; Akinfaderin, 2020; Van Biljon et al., 2020; Dossou & Emezue, 2020; Tapo et al., 2020; Lakew et al., 2020; Duh et al., 2020; Martinus et al., 2020; Hacheme, 2021; Ahia et al., 2021; Agyei et al., 2021; Emezue & Dossou, 2021; Reid et al., 2021).

To test the translation capabilities and effectiveness of the WON dataset, multiple NMT models were trained and fine-tuned: Existing bilingual models, pre-trained on related Bantu languages (isiZulu and kiSwahili) were selected to boost translation performance for the small WON corpus (Nyoni & Bassett, 2021). In addition, we trained a model on the constitutional dataset and fine-tuned it on WON to determine the potential for transfer across domains and dialects. Furthermore, inspired by the promising results for fine-tuning multilingual NMT models on little data of previously unseen languages (Adelani et al., 2022), we fine-tuned the M2M-100 (Fan et al., 2021) model on WON. The M2M-100 was trained on parallel data obtained from web crawls to translate between pairs of 100 languages, including several Bantu languages.

Data The WON dataset is processed by removing duplicate sentences and shuffled to remove creation order bias. The test set is then extracted, before the remaining corpus is combined (in ratio 4:1) with the Q&A data, to form the train and validation sets. The overall data splits are defined as 70% training, 10% validation, and 20% test set. The Q&A data is directly combined with the WON data for training, since it is written in the same dialect, and closer in domain and length than the constitution data. We refer to this combined data as WON+.

Models The transformer architecture (Vaswani et al., 2017) is used for the bilingual translation models and is implemented using JoeyNMT (Kreutzer et al., 2019). The architecture and hyperparameters are the same for all bilingual models to ensure comparability of results. It has 6 encoder and 6 decoder layers, 4 attention heads, and an embedding size of 256, hidden size of 1024, and is trained with Adam (Kingma & Ba, 2015) and batches of 4096 tokens. Byte Pair Encoding (Sennrich et al., 2016) is used to create a shared vocabulary of 4000 tokens. The learning (0.0001–0.0003) and dropout (0.05–0.3) rate are tuned for the fine-tuning stages of the pre-trained models. The models for kiSwahili and isiZulu had been trained on JW300 (Agić & Vulić, 2019) and were obtained from the

Masakhane-MT repository (∀ et al., 2020).[9] They scored 48.79 BLEU and 38.33 BLEU respectively on JW300 test sets.

Results Table 2 presents the translation quality of the above models on the WON test data. By training on WON+ from scratch, we reach a respectable 8.9 BLEU. When fine-tuning the other bilingual transformer models on this data, translation quality is further improved for the transfer from kiSwahili (+8.7) and isiZulu (+5.9). This strategy is surprisingly effective, since the models score <1 BLEU before fine-tuning (zero-shot testing). Transfer from the constitution translation is disappointingly unsuccessful (worse than training from scratch), probably because of the domain gap. The overall best results are obtained by fine-tuning the massively multilingual M2M-100 model, reaching 24.2 BLEU. The unanimous increase in BLEU scores after fine-tuning the pre-trained models show that the WON created is very effective for machine translation when combined with transfer learning. Translation examples are shown in Appendix A.4. We are currently collecting impressions of their quality and robustness by the workshop participants.

5 Conclusion, Future Work and Impact

We conducted a data creation workshop for Oshiwambo speakers designed to represent their culture and to support the language development across generations. The collected translations made it possible to train a first NMT system for Oshindonga with respectable quality. This demonstrated how a relatively small but well-designed and targeted effort can go a long way for languages that have previously been excluded from language technologies. We outline the future of the NMT development and the data use below, and we discuss the project's larger impact for language conservancy.

5.1 The Future of the Collected Data

While we are writing this paper, the post-processing of the data is still ongoing. We are in the process of defining under which conditions the data may be shared with other researchers or communities. The participants of the data creation workshop are *co-authors and co-owners* of the resulting data, hence they should receive remuneration for any future commercial or external use, in proportion of how many sentences they contributed. Furthermore, it should be in their power to decide for which purposes the data is used (Paullada, 2020).

5.2 The Future of Oshiwambo NMT Development

The NMT model developed here serves as a proof-of-concept for obtaining funding and finding more stakeholders and participants. As analyzed in Section 4.1, the collected data is relatively simple and does not contain long or complex sentences, as the generation of the participants has a diluted knowledge of Oshiwambo. One desirable use case for the deployment of an NMT model would be in negotiations between English and Oshiwambo speakers. Contracts are often defined in English, and Oshiwambo speakers who are not fluent in English rely on middlemen to interpret, who might be corrupt, manipulative and unfaithful. Ideally, contracts would be written in Oshiwambo first, so that Oshiwambo speakers can define their own terms, and then they would get translated into English with an NMT system. In order to train an NMT system for this scenario, more complex in-domain data has to be recorded and translated, in collaboration with Oshiwambo elderly who have a better notion of the language and the ability to articulate entire dialogues and complex matters.

5.3 Impact: Towards Language Conservancies

Our data creation initiative is a first step towards providing a sustainable income to communities across generations through cultural heritage conservancies, as opposed to the popular wildlife recovery conservancies. Namibia today has 86 communal conservancies, which are "self-governing, democratic entities, run by their members, with fixed boundaries that are agreed with adjacent conservancies, communities or land owners".[10] They span tourism and trophy hunting especially in areas with an abundance of wildlife, but on marginal lands, the cultural heritage is in the foreground.

[9] https://github.com/masakhane-io/masakhane-mt

[10] https://conservationtourism.com.na/communal-conservancies

Even though tourism can provide income, most rural communities and conservancy members rely on farming as their main source of income, which is now heavily threatened by climate change. Alternative streams of income become increasingly more important, especially as 52% of Namibia's population (in particular the younger generations) is currently living in urban areas, drawn there to find jobs, but often struggling with high living and tertiary education costs. Our vision is that the collaborative process of data creation, as exemplified in this paper, can provide an alternative, *self-determined and sustainable stream of income*, that simultaneously produces cultural datasets reflecting ancient traditions and processes, and preserves knowledge for future generations to mitigate climate-change related challenges, and creates opportunities for younger generations to develop indigenous-inspired and community-serving technologies and NLP tools.

ACKNOWLEDGEMENTS

We would firstly like to thank the main sponsors of the research, namely Google, Jungwon Byun and NamibiaNLP Crowdfunding donors, with the special assistance of Erin LeDell and Davor Orlc. Many thanks to all the reviewers as well as the anonymous AfricaNLP reviewers for their valuable feedback. Further, we thank Tosin Adewumi for the permission to use the Q&A data, and David Adelani for providing the M2M fine-tuning code. Additionally, special mention are given to Rosalia, Tuli, Tulela and Tulonga Kapiye for their invaluable continuous motivation and support. A full acknowledgement list can be found in the appendix A.6.

REFERENCES

Ahmed Abdelali, Francisco Guzman, Hassan Sajjad, and Stephan Vogel. The AMARA corpus: Building parallel language resources for the educational domain. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 1856–1862, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `http://www.lrec-conf.org/proceedings/lrec2014/pdf/877_Paper.pdf`.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen Hassan Muhammad, Guyo Dub Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles HACHEME, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ayoade Ajibade, Tunde Oluwaseyi Ajayi, Yvonne Wambui Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Koffi KALIPE, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, and Ayodele Awokoya. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, 2022.

Tosin Adewumi. Ìtàkúròso: Exploiting cross-lingual transferability for natural language generation of dialogues in low-resource, african languages. *AfricaNLP Workshop*, 2022. URL `https://openreview.net/forum?id=BtZlF5M4Lb9`.

Željko Agić and Ivan Vulić. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3204–3210, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1310. URL `https://aclanthology.org/P19-1310`.

Emmanuel Agyei, Xiaoling Zhang, Sophyani Banaamwini Yussif, and Bless Lord Y Agbley. Akan-english: Transformer for low resource translation. In *2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 256–259. IEEE, 2021.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3316–3333, Punta Cana, Dominican Republic, November

2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.282. URL https://aclanthology.org/2021.findings-emnlp.282.

Adewale Akinfaderin. HausaMT v1.0: Towards English–Hausa neural machine translation. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pp. 144–147, Seattle, USA, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.winlp-1.38. URL https://aclanthology.org/2020.winlp-1.38.

D Ashton, Tangeni Iijambo, M Matengu, and E Kalenga. Implementation of the government of the republic of namibia's language policy for schools in selected primary schools in windhoek. 2008.

Steven Bird. Last words: Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474, September 2009. doi: 10.1162/coli.35.3.469. URL https://aclanthology.org/J09-3007.

Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3504–3519, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.313. URL https://aclanthology.org/2020.coling-main.313.

Christo Botha. The odendaal plan: "development" for colonial namibia. 2022. URL https://www.namibweb.com/oden.htm.

Andrea Carmantini, Peter Bell, and Steve Renals. Untranscribed web audio for low resource speech recognition. In *INTERSPEECH*, pp. 226–230, 2019.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.579. URL https://aclanthology.org/2020.coling-main.579.

Hilaria Cruz and Joseph Waring. Deploying technology to save endangered languages. *CoRR*, abs/1908.08971, 2019. URL http://arxiv.org/abs/1908.08971.

Bonaventure FP Dossou and Chris C Emezue. Ffr v1. 1: Fon-french neural machine translation. *arXiv preprint arXiv:2006.09217*, 2020.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14757–14765, 2021.

Kevin Duh, Paul McNamee, Matt Post, and Brian Thompson. Benchmarking neural and statistical machine translation on low-resource African languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 2667–2675, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.325.

Chris Chinenye Emezue and Bonaventure F. P. Dossou. MMTAfrica: Multilingual machine translation for African languages. In *Proceedings of the Sixth Conference on Machine Translation*, pp. 398–411, Online, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.wmt-1.48.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48, 2021.

∀, Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi

Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2144–2160, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.195. URL https://aclanthology.org/2020.findings-emnlp.195.

Niklaas Johannes Fredericks. Challenges facing the development of namibian languages. a Conference of Harmonization of South African languages, 2007.

Jenna Frydman. A critical analysis of namibia's english-only language policy. *Selected Proceedings of the 40th Annual Conference on African Linguistics*, 01 2011.

Matthew Anthony Gallo. Bantu education, and its living educational and socioeconomic legacy in apartheid and post-apartheid south africa. Master's thesis, 2020.

Gilles Hacheme. English2gbe: A multilingual machine translation model for {Fon/Ewe} gbe. *arXiv preprint arXiv:2112.11482*, 2021.

Riikka Halme. *A tonal grammar of Kwanyama*, volume 8. R. Köppe, 2004.

Mika Hämäläinen. Endangered languages are not low-resourced! In *Multilingual Facilitation*, pp. 1–11. University of Helsinki, 2021.

Paavo Hasheela. *Omishe di dule eyovi*. Gamsberg Macmillan, 1993.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Julia Kreutzer, Jasmijn Bastings, and Stefan Riezler. Joey NMT: A minimalist NMT toolkit for novices. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pp. 109–114, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-3019. URL https://aclanthology.org/D19-3019.

Surafel Melaku Lakew, Matteo Negri, and Marco Turchi. Low resource neural machine translation: A benchmark for five african languages. 2020.

Laura Martinus and Jade Z. Abbott. A focus on neural machine translation for african languages. *CoRR*, abs/1906.05685, 2019. URL http://arxiv.org/abs/1906.05685.

Laura Martinus, Jason Webster, Joanne Moonsamy, Moses Shaba Jnr, Ridha Moosa, and Robert Fairon. Neural machine translation for south africa's official languages. *arXiv preprint arXiv:2005.06609*, 2020.

Ruth Morgan, Scott Liddell, Marius Haikali, Sackeus P Ashipala, Polo Daniel, Hilifilua ET Haiduwah, Rauna Ndeshihafela Hashiyana, Nangolo Jeremia Israel, Festus Tshikuku Linus, Henock Hango Niilenge, et al. Namibian sign language to english and oshiwambo. 1991.

Ngila R. L. Mwase. The repatriation, rehabilitation and resettlement of namibian refugees at independence. *Community Development Journal*, 25(2):113–121, 1990. ISSN 00103802, 14682656. URL http://www.jstor.org/stable/44256860.

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, et al. A summary of the first workshop on language technology for language documentation and revitalization. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, pp. 342, 2020.

Martin Ngodji. The story of the bible among ovakwanyama: the agency of indigenous translators. Master's thesis, 2004.

Evander Nyoni and Bruce A Bassett. Low-resource neural machine translation for southern african languages. *AfricaNLP Workshop*, 2021.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pp. 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.11. URL `https://aclanthology.org/2021.mrl-1.11`.

Amandalynne Paullada. How does machine translation shift power? *Resistance in AI Workshop*, 2020. URL `https://drive.google.com/file/d/1wO5UOxTThrcCiU-gEJm_KBShxL_YXEXx/view`.

Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL `https://aclanthology.org/W18-6319`.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1306–1320, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.99. URL `https://aclanthology.org/2021.emnlp-main.99`.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1162. URL `https://aclanthology.org/P16-1162`.

Immanuel N Shatepa and Edward T Shikesho. The case of oshiwambo-english/english-oshiwambo translation of informative texts: Is meaning lost in translation? *JULACE: Journal of the University of Namibia Language Centre*, 4(1):70–79, 2019.

Talita C. Smit. Is 'english-centric bilingualism' suffocating namibian national and indigenous languages? In *Journal of Language and Communication*, volume 6, 2012. URL `https://link.gale.com/apps/doc/A341458397/LitRC?u=anon~d091f1b4&amp;sid=googleScholar&amp;xid=c0c0a37b`.

Gerald Stell. *English in Namibia. A socio-historical account.*, pp. 22–41. 04 2021. ISBN 9789027209191. doi: 10.1075/veaw.g65.02ste.

SWAPO. *Their Blood Waters Our Freedom: Glory to the Heroes and Heroines of the Namibian Liberation Struggle*. SWAPO Party, 1996. ISBN 9789991673806. URL `https://books.google.com.na/books?id=1ml0QgAACAAJ`.

Allahsera Auguste Tapo, Bakary Coulibaly, Sébastien Diarra, Christopher Homan, Julia Kreutzer, Sarah Luger, Arthur Nagashima, Marcos Zampieri, and Michael Leventhal. Neural machine translation for extremely low-resource African languages: A case study on Bambara. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pp. 23–32, Suzhou, China, December 2020. Association for Computational Linguistics. URL `https://aclanthology.org/2020.loresmt-1.3`.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

UNESCO. In *International Expert Meeting on the UNESCO Programme "Safeguarding of Endangered Languages"*, 2003.

Trusteeship United Nations Department of Political Affairs and Decolonization. Decolonization — issue on namibia. 1977. URL https://www.un.org/dppa/decolonization/sites/www.un.org.dppa.decolonization/files/decon_num_9-1.pdf.

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. On optimal transformer depth for low-resource language translation. 2020.

Daan van Esch, Ben Foley, and Nay San. Future directions in technological support for language documentation. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pp. 14–22, Honolulu, February 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-6003.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 3423–3430. European Language Resources Association (ELRA), 2010.

Heike Winschiers-Theophilus, Shilumbe Kuria, Gereon Kapuire, Nicola Bidwell, and Edwin Blake. Being participated - a community approach. pp. 1–10, 11 2010. doi: 10.1145/1900441.1900443.

Anssi Yli-Jyrä. Explorations on positionwise flag diacritics in finite-state morphology. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pp. 262–269, Riga, Latvia, May 2011. Northern European Association for Language Technology (NEALT). URL https://aclanthology.org/W11-4636.

Anssi Yli-Jyrä. On finite-state tonology with autosegmental representations. In *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing*, pp. 90–98, St Andrews, Scotland, July 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-1814.

Anssi Yli-Jyrä. Three equivalent codes for autosegmental representations. In *Proceedings of the 12th International Conference on Finite-State Methods and Natural Language Processing 2015 (FSMNLP 2015 Düsseldorf)*. Association for Computational Linguistics, 2015. URL https://aclanthology.org/W15-4812.

A Appendix

A.1 Alignment of the Constitution

We obtain the Constitution of Namibia in English and Oshikwanyama from `https://www.kas.de/`. The respective files are formatted consistently, and thereby can easily be parsed with `pdftotext`.[11] The largest hindrance for automatic alignment is that the English version (2018) contains annotations and more recent amendments that are not present in the Oshiwambo version (2017). We discard these, and resolve the corresponding alignment shifts manually. We only use the sentences that are part of the 148 articles of the Constitution, excluding the table of contents, the preface and the schedules.

A.2 Workshop Schedule

Day	Topic	In-class?	Supervised?	N
before	interviews, warm-up, familiarization with translation			
1	names, seed words	x	x	6
2	wildlife, agriculture	(x)	(x)	6
3	topics of participants' interest	x		6
4	*rest*			
5	culture, "ask a veteran", climate change	x	x	8
6	oral comprehension (radio news)	x	x	8
7	human-environment relationship (totems), economic systems, leadership, societal challenges	x	x	8
8	free writing, profession-based Q&A	x	(x)	4
after	revising, completing translations			

Table 3: The schedule per day: topics for data creation, the method of supervision ("(x)"=partially) and instruction and the number of participants (N).

A.3 Seed Words

Figure 2 shows two collections of seed keywords to start the sentence creation, the board of the classroom after a supervised session, and a digital list of keywords extracted in the unsupervised radio comprehension task.

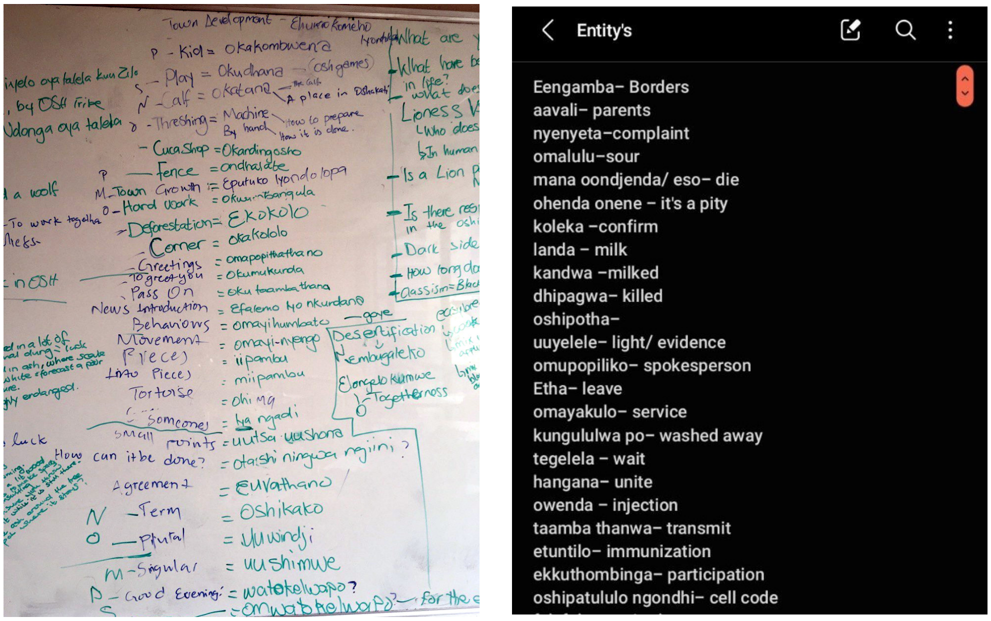
A.4 NMT Translation Examples

Table 4 shows generated translations by the M2M model before and after fine-tuning, for sentences from the test set. It can be seen that while the pre-trained predictions are coherent (due to the large amount of data used to train the m2m model), the translation accuracy increases drastically after fine-tuning, in line with the quantitative BLEU scores received.

A.5 Reflections on the Data Creation Process

The daily sentence target number has been smaller than expected, mainly due to the discussion and debates on how to write certain words and what the accurate translation would be. Future data collections can be started from the same seed words and their now established contexts, such that the discussion phase can be shortened. We expect no risk of sentence duplicates, since every word allows for a rich variety of use cases and personal perspectives to be reflected in the sentences that different participants would generate.

[11] `https://linux.die.net/man/1/pdftotext`



Eengamba- Borders
aavali- parents
nyenyeta-complaint
omalulu-sour
mana oondjenda/ eso- die
ohenda onene - it's a pity
koleka -confirm
landa - milk
kandwa -milked
dhipagwa- killed
oshipotha-
uuyelele- light/ evidence
omupopiliko- spokesperson
Etha- leave
omayakulo- service
kungululwa po- washed away
tegelela - wait
hangana- unite
owenda - injection
taamba thanwa- transmit
etuntilo- immunization
ekkuthombinga- participation
oshipatululo ngondhi- cell code

(a) Board after a classroom brainstorming session.

(b) Radio word extraction home exercise.

Figure 2: Examples of keyword collections in Oshiwambo.

Description	Sentence
Source (ng)	Ngele oho simaneke aankuluntu oto ka la wuna omwenyo omule.
Reference (en)	If you respect elders you will have a long life.
Prediction before fine-tuning (en)	Thou shalt speak to the wicked, thou shalt speak to the wicked.
Prediction after fine-tuning (en)	If you respect parents you will have a long life.
Source (ng)	Aantu otaya taamba iikulya yoshukukuta kombelewa
Reference (en)	People are receiving drought food at the office
Prediction before fine-tuning (en)	I will be able to do it, and I will be able to do it.
Prediction after fine-tuning (en)	People are collecting cooking food in the office
Source (ng)	Ngele o wa hala o ku koka owuna o ku lya onyama oyindji.
Reference (en)	If you want to grow you have to eat meat.
Prediction before fine-tuning (en)	It is still in the midst of it, and it is still in the midst of it.
Prediction after fine-tuning (en)	If you want to grow you need to eat a lot of meat.

Table 4: Translation examples for sentences from the test set, generated by the M2M model before and after fine-tuning.

A.6 Full List of Acknowledgements

The success of the ongoing work in creating Oshiwambo to English translations, is largely thanks to the following individuals and organisations for their support, financially and in kind.

Participants: Otto Natanial, Nelson Nelumbuh, Mekondjo Alfeus, Tulimeke Nghipopi, Taliya Luuli, Selma, Paulus Mwafuda, Loide, Aaron Nangobe and Maria.

Main Sponsors: Google, Jungwon Byun and NamibiaNLP Crowdfunding donors.

Fiscal Sponsors: Women in Machine Learning and Data Science (Erin LeDell) and Knowledge 4 All (Davor Orlc).

Other Sponsors: BBC Cafe Windhoek, Gondwana Collection, Franco Namibia Cultural Centre, Cafe Prestige, Droombos and Vineyard Estate.

NLP Paper Reviewers: Jay Loftstead, Margareth Amagulu, Kristina Georgiva Young, Luisa Teixera, Daan Van Esch, Martha Kamkhuema and the AfricaNLP reviewers.

Data and Models: Tosin Adewumi (Q&A data) and David Adelani (M2M fine-tuning code).

Cultural Advisors: Mr Nahas Angula, Mr Vinnia Ndadi, Margareth Amagulu, Colonel Theophilia Shaanika (retired), Brigadier General Fiina Amupolo (retired), Josua and Helvi Nekoto.

Special Mention: Masakhane, Deep Learning Indaba, Dot.Astronomy, SeafoodBroil Namibia, Rosalia, Tuli, Tulela and Tulonga Kapiye.