LREC-COLING 2024

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL)

Workshop Proceedings

Editors
Rooweither Mabuya, Muzi Matfunjwa, Mmasibidi Setaka,
and Menno van Zaanen

25 May, 2024
Torino, Italia

Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL)

ISBN 978-2-493814-40-1
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

Africa is a multilingual continent with an estimation of 1500 to 2000 indigenous languages. Many of the languages currently have no or very limited language resources available and are often structurally quite different from more well-resourced languages, therefore requiring the development and use of specialized techniques. To bring together and emphasize research in these areas, the Resources for African Indigenous Languages (RAIL) workshop series aims to provide an interdisciplinary platform for researchers working on resources (data collections, tools, etc.) specifically targeted towards African indigenous languages. These events provide an overview of the current state-of-the-art and emphasize the availability of African indigenous language resources, including both data and tools.

With the UNESCO-supported Decade of Indigenous Languages, there is currently much interest in indigenous languages. The Permanent Forum on Indigenous Issues mentioned that "40 percent of the estimated 6,700 languages spoken around the world were in danger of disappearing" and the "languages represent complex systems of knowledge and communication and should be recognized as a strategic national resource for development, peace building and reconciliation."

This year's RAIL workshop is the fifth in the series. The first RAIL workshop was co-located with the Language Resources and Evaluation Conference (LREC) in 2020, whereas the second RAIL workshop in 2021 was co-located with the Digital Humanities Association of Southern Africa (DHASA) conference. Both events were virtual. The third RAIL workshop was co-located with the tenth Southern African Microlinguistics Workshop (SAMWOP) and took place in person in 2022 in Potchefstroom, South Africa. The fourth RAIL workshop was co-located with the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL) in Dubrovnik, Croatia in 2023.

Previous RAIL workshops showed that the presented problems (and solutions) are typically not only applicable to African languages. Many issues are also relevant to other low-resource languages, such as different scripts and properties like tone. As such, these languages share similar challenges. This allows for researchers working on these languages with such properties (including non-African languages) to learn from each other, especially on issues about language resource development.

For the fifth RAIL workshop, in total, 39 high-quality submissions were received. Out of these, 17 submissions (15 long papers and 2 short papers) were selected for presentation in the workshop. All submissions received three reviews using a double-blind review process. This RAIL workshop took place as a full day workshop in Lingotto Conference Centre, Torino, Italy on 25 May 2024. It was co-located with the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). Each presentation consisted of 25 minutes for long papers (including time for discussion) and 10 minutes for short papers.

This publication adheres to South Africa's DHET's 60% rule, authors in the proceedings come from a wide range of institutions.

This RAIL workshop's theme was "Creating resources for less-resourced languages", but submissions on any topic related to properties of African indigenous languages were considered. Several suggested topics for the workshop were mentioned in the call for papers:

- Digital representations of linguistic structures;
- Descriptions of corpora or other data sets of African indigenous languages;
- Building resources for (under-resourced) African indigenous languages;
- Developing and using African indigenous languages in the digital age;
- Effectiveness of digital technologies for the development of African indigenous languages;
- Revealing unknown or unpublished existing resources for African indigenous languages;
- Developing desired resources for African indigenous languages;
- Improving quality, availability and accessibility of African indigenous language resources.

The goals for the workshop were:

- to bring together researchers who are interested in showcasing their research and thereby boosting the field of African indigenous languages,
- to create the conditions for the emergence of a scientific community of practice that focuses on data, as well as tools, specifically designed for or applied to indigenous languages found in Africa,
- to create conversations between academics and researchers in different fields such as African indigenous languages, computational linguistics, sociolinguistics, and language technology, and
- to provide an opportunity for the African indigenous languages community to identify, describe and share their language resources.

We would like to mention explicitly that the term “indigenous languages” used in the RAIL workshop is intended to refer to non-colonial languages (in this case those used in Africa). In no way is this term used to cause any harm or discomfort to anyone. Many of these languages were or still are marginalized and the workshop aims to bring attention to the creation, curation, and development of resources for these languages in Africa.

The organizers would like to thank the authors who submitted manuscripts and the programme committee who provided feedback on the quality and content of the submissions.

The RAIL organizing committee and editors of the proceedings

- Rooweither Mabuya, South African Centre for Digital Language Resources
- Muzi Matfunjwa, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Menno van Zaanen, South African Centre for Digital Language Resources

Organizing Committee

- Rooweither Mabuya, South African Centre for Digital Language Resources
- Muzi Matfunjwa, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Menno van Zaanen, South African Centre for Digital Language Resources

Programme Committee

- Andiswa Bukula, South African Centre for Digital Language Resources
- Ayodele Akinola, Chrisland University
- Benito Trollip, South African Centre for Digital Language Resources
- Deon du Plessis, South African Centre for Digital Language Resources
- Elias Malete, University of the Free State
- Elsabe Taljard, University of Pretoria
- Emmanuel Ngue Um, University of Yaoundé I
- Febe De Wet, Stellenbosch University
- Friedel Wolff, South African Centre for Digital Language Resources
- Heather Brookes, University of Stellenbosch
- Hussein Suleman, University of Cape Town
- Innocentia Mhlambi, University of the Witwatersrand
- Johannes Sibeko, Nelson Mandela University
- Juan Steyn, South African Centre for Digital Language Resources
- Kaka Mokakale, North-West University
- Laurette Marais, Council for Scientific and Industrial Research
- Malefu Mahloane, University of the Free State
- Maria Keet, University of Cape Town
- Marissa Griesel, University of South Africa
- Martin Puttkammer, North-West University
- Menno van Zaanen, South African Centre for Digital Language Resources
- Mmasibidi Setaka, South African Centre for Digital Language Resources
- Mpho Raborife, University of Johannesburg
- Muzi Matfunjwa, South African Centre for Digital Language Resources

- Nomsebenzi Malele, University of South Africa
- Nuette Heyns, North-West University
- Papi Lemeko, Central University of Technology
- Pule Phindane, Central University of Technology
- Roald Eiselen, North-West University
- Rooweither Mabuya, South African Centre for Digital Language Resources
- Sibonelo Dlamini, University of KwaZulu-Natal
- Tanja Gaustad, North-West University
- Temitope Kekere, University of Pretoria
- Tunde Ope-Davies, University of Lagos

Table of Contents

Doing Phonetics in the Rift Valley: Sound Systems of Maasai, Iraqw and Hadza
Alain Ghio, Didier Demolin, Michael Karani and Yohann Meynadier 1

Kallaama: A Transcribed Speech Dataset about Agriculture in the Three Most Widely Spoken Languages in Senegal
Elodie Gauthier, Aminata Ndiaye and Abdoulaye Guissé 10

Long-Form Recordings to Study Children's Language Input and Output in Under-Resourced Contexts
Joseph R. Coffey and Alejandrina Cristia 20

Developing Bilingual English-Setswana Datasets for Space Domain
Tebatso G. Moape, Sunday Olusegun Ojo and Oludayo O. Olugbara 32

Compiling a List of Frequently Used Setswana Words for Developing Readability Measures
Johannes Sibeko 37

A Qualitative Inquiry into the South African Language Identifier's Performance on YouTube Comments.
Nkazimlo N. Ngcungca, Johannes Sibeko and Sharon Rudman 45

The First Universal Dependency Treebank for Tswana: Tswana-Popapolelo
Tanja Gaustad, Ansu Berg, Rigardt Pretorius and Roald Eiselen 55

Adapting Nine Traditional Text Readability Measures into Sesotho
Johannes Sibeko and Menno van Zaanen 66

Bootstrapping Syntactic Resources from isiZulu to Siswati
Laurette Marais, Laurette Pretorius and Lionel Clive Posthumus 77

Early Child Language Resources and Corpora Developed in Nine African Languages by the SADiLaR Child Language Development Node
Michelle J. White, Frenette Southwood and Sefela Londiwe Yalala 86

Morphological Synthesizer for Ge'ez Language: Addressing Morphological Complexity and Resource Limitations
Gebrearegawi Gebremariam Gidey, Hailay Kidu Teklehaymanot and Gebregewergs Mezgebe Atsbha 94

EthioMT: Parallel Corpus for Low-resource Ethiopian Languages
Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh and Jugal Kalita 107

Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme
Nuhu Ibrahim, Felicity Mulford, Matt Lawrence and Riza Batista-Navarro 115

Low Resource Question Answering: An Amharic Benchmarking Dataset
Tilahun Abedissa Taffa, Ricardo Usbeck and Yaregal Assabie 124

The Annotators Agree to Not Agree on the Fine-grained Annotation of Hate-speech against Women in Algerian Dialect Comments
Imane Guellil, Yousra Houichi, Sara Chennoufi, Mohamed Boubred, Anfal Yousra Boucetta and Faical Azouaou 133

Advancing Language Diversity and Inclusion: Towards a Neural Network-based Spell Checker and Correction for Wolof
Thierno Ibrahima Cissé and Fatiha Sadat 140

Lateral Inversions, Word Form/Order, Unnamed Grammatical Entities and Ambiguities in the Constituency Parsing and Annotation of the Igala Syntax through the English Language
Mahmud Mohammed Momoh 152

Workshop Program

Saturday, May 25, 2024

09:00–09:05 ***Opening***

09:05–09:30 *Doing Phonetics in the Rift Valley: Sound Systems of Maasai, Iraqw and Hadza*
Alain Ghio, Didier Demolin, Michael Karani and Yohann Meynadier

09:30–09:55 *Kallaama: A Transcribed Speech Dataset about Agriculture in the Three Most Widely Spoken Languages in Senegal*
Elodie Gauthier, Aminata Ndiaye and Abdoulaye Guissé

09:55–10:20 *Long-Form Recordings to Study Children's Language Input and Output in Under-Resourced Contexts*
Joseph R. Coffey and Alejandrina Cristia

10:20–10:30 *Developing Bilingual English-Setswana Datasets for Space Domain*
Tebatso G. Moape, Sunday Olusegun Ojo and Oludayo O. Olugbara

10:30–11:00 ***Coffee break***

11:00–11:25 *Compiling a List of Frequently Used Setswana Words for Developing Readability Measures*
Johannes Sibeko

11:25–11:50 *A Qualitative Inquiry into the South African Language Identifier's Performance on YouTube Comments.*
Nkazimlo N. Ngcungca, Johannes Sibeko and Sharon Rudman

11:50–12:15 *The First Universal Dependency Treebank for Tswana: Tswana-Popapolelo*
Tanja Gaustad, Ansu Berg, Rigardt Pretorius and Roald Eiselen

12:15–12:40 *Adapting Nine Traditional Text Readability Measures into Sesotho*
Johannes Sibeko and Menno van Zaanen

12:40–13:05 *Bootstrapping Syntactic Resources from isiZulu to Siswati*
Laurette Marais, Laurette Pretorius and Lionel Clive Posthumus

13:05–14:20 ***Lunch break***

Saturday, May 25, 2024 (continued)

14:20–14:45 *Early Child Language Resources and Corpora Developed in Nine African Languages by the SADiLaR Child Language Development Node*
Michelle J. White, Frenette Southwood and Sefela Londiwe Yalala

14:45–15:10 *Morphological Synthesizer for Ge'ez Language: Addressing Morphological Complexity and Resource Limitations*
Gebrearegawi Gebremariam Gidey, Hailay Kidu Teklehaymanot and Gebregewergs Mezgebe Atsbha

15:10–15:35 *EthioMT: Parallel Corpus for Low-resource Ethiopian Languages*
Atnafu Lambebo Tonja, Olga Kolesnikova, Alexander Gelbukh and Jugal Kalita

15:35–16:00 *Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme*
Nuhu Ibrahim, Felicity Mulford, Matt Lawrence and Riza Batista-Navarro

16:00–16:30 ***Coffee break***

16:30–16:55 *Low Resource Question Answering: An Amharic Benchmarking Dataset*
Tilahun Abedissa Taffa, Ricardo Usbeck and Yaregal Assabie

16:55–17:05 *The Annotators Agree to Not Agree on the Fine-grained Annotation of Hatespeech against Women in Algerian Dialect Comments*
Imane Guellil, Yousra Houichi, Sara Chennoufi, Mohamed Boubred, Anfal Yousra Boucetta and Faical Azouaou

17:05–17:30 *Advancing Language Diversity and Inclusion: Towards a Neural Network-based Spell Checker and Correction for Wolof*
Thierno Ibrahima Cissé and Fatiha Sadat

17:30–17:55 *Lateral Inversions, Word Form/Order, Unnamed Grammatical Entities and Ambiguities in the Constituency Parsing and Annotation of the Igala Syntax through the English Language*
Mahmud Mohammed Momoh

17:55–18:00 ***Closing***

Doing Phonetics in the Rift Valley : Sound Systems of Maasai, Iraqw and Hadza

Ghio[1] A., Demolin[2] D., Karani[3] M., Meynadier[1] Y.

(1) Aix Marseille Univ, CNRS, LPL, Aix-en-Provence, France
(2) LPP (UMR7018, CNRS/Sorbonne Nouvelle), Paris, France
(3) Centre for Communication Studies, College of Humanities, Univ of Dar es Salaam, Tanzania

{alain.ghio,yohann.meynadier}@univ-amu.fr,didier.demolin@sorbonne-nouvelle.fr, karanim@udsm.ac.tz

Abstract

This article discusses the contribution of experimental techniques to the recording of phonetic data in the field. Only a small part of the phonological systems of African languages is described with precision. This is why it is important to collect empirical data in the form of sound, video and physiological recordings. This allows research questions such as patterns of variation to be addressed. Analytical methods show how to interpret data from physical principles and integrate them into appropriate models. The question of linguistic contact between different language families is also addressed. To achieve these general objectives, we present the way we design corpora, and the different ways of recording data with crucial technical considerations during fieldwork. Finally, we focus on 3 languages spoken in the Great African Rift Zone, which includes several linguistic areas belonging to the four major linguistic families of the continent. (1) Hadza is a click language with a very complex consonant system. (2) Iraqw is a Cushitic language with ejective consonants. (3) Maasai is a Nilotic language with implosive consonants and a very elaborate set of interjections, ideophones and animal calls that include sounds not described in the International Phonetic Alphabet.

Keywords: experimental phonetics, fieldwork, hadza, iraqw, maasai

1. The need for empirical data

More than 2000 languages are spoken on the African continent (Güldemann, 2018). Only a small part of them has been described in grammar form. Among those that have been described, the part of grammar devoted to phonetics and phonology is often less than 10% of the content of the full grammar (Maddieson, 2002). In addition, Maddieson indicated that “while syntactic patterns are documented with example sentences, often from natural discourse or texts, the phonetic facts are rarely if ever documented by the presentation of hard evidence”. The phonological descriptions to which we have access are sometimes imprecise and there is often an ambiguity of symbols used. To illustrate the lack of data, we can compare the map of languages described in the PHOIBLE database (Moran, S., & McCloy, 2019) which contains 2186 distinct languages and the actual map of languages spoken in Tanzania (Figure 1). Although this database is the most extensive in the field, only 30 languages are proposed in PHOIBLE among the 125 ones spoken in Tanzania. Not only is the data reduced, but it also contains inaccuracies and even errors. For example, if you select the Iraqw Language (Glottocode: iraq1241), the Iraqw sound inventory proposed by the Stanford Phonology Archive (SPA) or by the UCLA Phonological Segment Inventory Database (UPSID) includes implosive consonants in this language as [ɓ] and [ɗ], which is false considering the detailed work by Mous (1993). Furthermore, still in the Iraqw Language, the consonant /q/ is most of the time described as a voiceless uvular plosive but the realization of this phoneme appears not as pulmonic but with an ejective mechanism, which would merit a phonological representation /q’/. In the same way, if you select the Maasai Language (Glottocode: masa1300), the Maasai sound inventory proposed by the Stanford Phonology Archive (SPA) or by the UCLA Phonological Segment Inventory Database (UPSID) includes a voiceless alveolar trill [r̥] whose voiceless feature is contested (Karani et al., 2023).

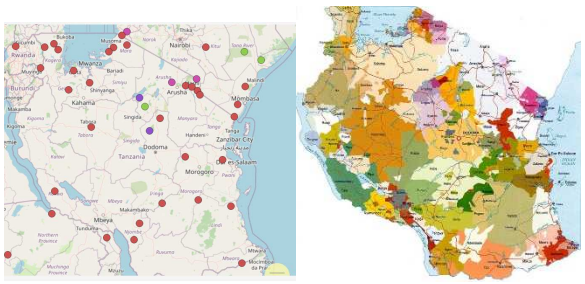


Figure 1: Languages spoken in Tanzania.
On the left: Inventory in PHOIBLE database (Moran, S., & McCloy, 2019).
On the right: the Atlasi ya lugha za Tanzania, (2009)

For all these reasons, it is important to collect empirical data in the form of sound, video and physiological recordings. This is the goal of the projects, “SYSORI” (Sound Systems of Rift Valley languages) and its extension “COSYSORI”, funded by CNRS (see §.6 acknowledgements). The objective of this article is to provide some preliminary details.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 1–9
25 May, 2024.

2. Elements of the methodology

2.1 Field linguistic or fieldwork at home

There are several ways to do field linguistics. If the grammar of the language, and in particular the phonological system, is precisely described, it is possible to use an ecological approach as recommended by Gasquet (2015). As we work on under documented languages, it is not possible for us to describe linguistic phenomena on the fly. We apply an approach based on the corpus.

We practice elicitation with speakers from a defined community, allowing the description of phonetic characteristics of the language studied. Field linguistics certainly requires specific preparation, knowledge of places, history, geography, adaptation to local culture, the use of interpreters, and contact with people. If we refer to Crowley's (2007) description, we practice “dirty feet” linguistics. Nothing to do with the recording of native speakers within universities, defined by (Crowley, 2007) as “fieldwork at home”. In the field, we have to face sociolinguistic realities (social hierarchies, tensions, discrimination, differences according to age, sex, etc.). We must also find solutions linked to practical, logistical and technical challenges. We can mention, for example, sometimes it is very difficult to access isolated villages in the countryside. We also think about the absence of electricity or an intermittent electricity supply, thus requiring the use of batteries and solar panels. We can cite the difficulties of obtaining a low-noise environment with the nearby presence of children, livestock, etc. Such practice requires good field experience and enough preparation before embarking on fieldwork missions.

2.2 Ethical aspects

In our fieldwork, an interpreter, in our case a native speaker who is a trained linguist, clearly explains to volunteers what the goal of the research project is. Participants are informed about the way research will be conducted as well as their rights in the project including freedom to withdraw themselves from the project without giving reasons to do so. As our participants do not always know how to read and/or write, this information and the participant's agreement is audio or video recorded. The consent recordings are kept in a special storage location shared by project team members because they are non-anonymized data. All scientific data is then anonymised through the use of codes. Only a local investigator keeps private information concerning the speakers in order to contact them again in the future if necessary for more fieldwork. Only general anonymous information (age, sex, languages they speak, and the level of education) is accessible as methodological data. To compensate for the time spent, all participants receive a small cash allowance, as do the people who coordinate the target speakers.

2.3 How do we design corpus?

For us, the main corpus is oriented towards phonetic description. To obtain relevant and usable data, our method is to create an ad hoc corpus to answer a particular question. Most of the time, it is a list of words which includes the phonemes which we think have phonetic qualities under investigation. For instance, to observe the contrast between modal vs labialised consonants in Iraqw, we proposed a list of words with the modal expressions (konki, dakaát, anága kií/, daktani…) vs labialized (lakwanti, múk tlakw, án aga kwandeékw, tatlkwa…). To study the non-pulmonic consonants, we select words with ejective consonants in Iraqw or implosive consonants in Maasai. In this type of study, it is important to contrast these units with pulmonic ones in order to observe the typicality of non-pulmonic contrasted with pulmonic ones. In all the cases, the lists should be well designed and balanced in terms of occurrences (same amount of data by group if we study a contrast). If possible, the target phoneme should be placed at the initial, median or final position. The left and right contexts of the target expression should be managed. It can vary if we consider that the context can influence the realization of the target phoneme.The way to elicit the corpus can vary, depending on the capabilities of the speakers. Reading the list is the easiest way but it means that the chosen words should be written without ambiguity and above all, the participants should be able to read fluently. This is far from being the case in the field. An alternative is to practice a repetition: a referent speaker pronounces the target and the participant repeats it. We can also reinforce the elicitation by the proposition translated into another language mastered by the participant. For instance, in Tanzania where the majority of people speak Kiswahili, we would start by the word in Kiswahili and then the word in the studied language. The amount of data must be large enough to obtain statistically significant results and above all to have legitimacy to generalize the results. It is also important to record at least 10 speakers (5 male & 5 female) in order to neutralize a possible atypical participant. It can be important to mix also young, medium and old people but most of the time, it is not possible to fulfil all the ideal conditions in the field since each field site has its unique settings and requirements due to its geography, socio-political and cultural aspects. During the recording of data, it is also better if the speakers repeat the utterance or corpus several times in order to see if the phonetic realization is stable or not.

The study of prosodic features needs another type of item. If the studied language has tones

where the variation of tone can change the meaning, it is possible to work with isolated minimal pairs: for instance in Maasai, áàdàmú ('he will remember me') vs áádámú ('I remember you'). This can be the first step to study the tonal system of the language. But for a substantial prosodic analysis, it is better to observe longer sequences; it can be sentences, texts or spontaneous speech. An valuable aspect of data collection is also to record spontaneous speech. It is important to collect stories and not only specific target data. The content can be a valuable element of heritage and further analysis of data to answer. In this situation, we try to record oral history using a video camera. The technical design is described further in the subsequent sections.

2.4 How do we record the corpus?

It is now possible to use applications, which can help with data recording. For instance, LIG-Aikuma is a mobile app for speech data collection and language documentation (Blachon et al., 2016). One particular capability is the "respeaking" mode where the operator can record first a referent speaker. After a short processing step, the experimenter can playback this reference for other speakers. This mode can avoid a fastidious repetition for the native referent speaker. It does not replace the physical presence of an interpreter or native speaker but it could allow us to record new data without the physical presence of the referent if we cannot do otherwise. This application has the advantage since it is portable and it can work on a mobile phone. Unfortunately, it does not work on a slightly more sophisticated station equipped with a laptop, a sound card and a good microphone. Indeed, in order to obtain good sound data, we prefer to use good electrostatic microphones connected to an external soundboard. For certain specific corpora, we also use specialized devices that allow us to record physiological data. We will describe this further.

2.5 Annotation and segmentation

Annotating a corpus consists of adding relevant information for its use. This consists of indicating orthographically the content of the oral data but also and above all (in phonetics) of precisely identifying the linguistic boundaries (segmentation). When the corpus is elicitated as we do, the transcription is trivial if the speaker pronounces correctly the list of items. But it is rarely perfect and we are faced with hesitation, errors, autocorrections… The second step is the phonetic segmentation where we need to set the contents and the boundaries of the spoken text. This process is time-consuming because, with no-documented languages, only a manual process is possible. We know that some projects propose some tools to automate this processing. We think for example the BULB project which aims at supporting the documentation of unwritten languages with the help of automatic speech and language processing, in particular automatic speech recognition (Adda et al., 2016). We also are aware of the CLD2025 project (https://anr.fr/Project-ANR-19-CE38-0015) whose goal is to facilitate the task of documenting endangered languages by leveraging the potential of computational methods. But for the moment, the use of these technologies is not yet possible due to the lack of data and knowledge of the languages studied.

2.6 Data sharing

In order to archive the data collected in the field, we asked ourselves the question of sharing data on cloud. In France, the Pangloss Collection is a digital library whose objective is to store and facilitate access to audio recordings in endangered languages of the world (Boyd et al., 2014). This platform is a potential final archive of the data we collect. For the moment, we mainly need a collaborative work platform in order to exchange raw data, enriched data, results or bibliography relating to our fieldwork between European and African partners. Our choice temporarily is the RESANA platform run by the French government that allows partners to perform the above-mentioned activities virtually.

3. Technical considerations

To collect our data, we have different technical designs depending on the purpose.

3.1 Simple audio design

It is very tempting to record speakers with a cell phone. This is the strategy used by the BULB project (Adda et al., 2016) using the application LIG-Aikuma (Blachon et al., 2016). This way is good for documenting a language. However, in a precise phonetic analysis, we need more controlled data.

Our minimal recording installation consists of a microphone connected to a sound card, which is connected to a laptop. We have 2 types of microphones: a professional head-worn condenser microphone (AKG C520) or a microphone on-stand (AKG C1000S). The advantage of the head-worn microphone is that it focuses on the speaker's speech and therefore considerably limits surrounding noise. It also maintains a constant distance which can allow measurements of variations in speech intensity. The downside of this equipment is that it must be attached to the speaker's head or ears, requiring touching the face or moving the hair, which is sometimes tricky. In addition, this type of poorly shielded microphone can be sensitive to electromagnetic interference leading to unwanted

electrical noise. This is not the case with a stand microphone, for example, AKG C 1000 S where its gold sputtered capsule housing makes the microphone extremely rugged against humidity or adverse condition. The position of the microphone is important. In order to avoid “pop” noises, which occur on plosive or fricative consonants, it is preferable to shift the microphone away from the speaker's axis and direct it towards the mouth, which forms an angle of approximately 45°.

The microphones that we use are condenser ones. They all need a phantom power because of condenser technologies. This is the first reason to use a sound card that directly powers the microphones via the cable. The second reason to use an external sound card is that the signal-to-noise ratio is better than recording directly with a laptop. The models of soundcard change quickly but we can give the models used at the moment: Focusrite Scarlet 2i2, Solid Static Logic SSL2, RME Fireface UCX.

It is also possible to use a portable wav/mp3 recorder, for example the famous Zoom H4n. We can use this device in a standalone mode because it integrates microphones, amplifiers, digitizer and storage. It is also possible to connect external microphones like the ones that we described above. This device can power condenser microphones. The drawback of this equipment is that it is not easy to control the recording level, especially if the device is set near the speaker and far from the operator. A possible incorrect level setting will only be discovered at the end of the recordings, which is very regrettable. The second drawback is that the recorded files are named automatically on the local storage and it is sometimes tricky to recover who and what was recorded on the labelled in the following format: File0035, File0078 etc.

3.2 Video Imaging design

As we mentioned in § 2.3, it is also important to record people telling stories. These recordings are rich data on heritage values. In this case, a video recording is preferred as a simple sound track.

The first simple way is to capture the images with a mobile phone or with a video camera where the sound is recorded with the integrated microphone. The problem with that is that most of the time, the camera is far from the speakers, the sound is imprecise, and the level of noise is high, especially if the scene is recorded outside.

In order to obtain good sound data, we use a setting with 2 external microphones. The main one is a Sennheiser MKH 416-P48, a shotgun interference tube microphone designed for film, radio, and television, especially for outside applications. This microphone has excellent directivity and a good sensitivity (25 mV/Pa). In order to reduce external noise, we equip it with a rigid windscreen MZW60-1 and we add above a windmuff MZH60-1. This microphone is set on a stand and the operator can manually orient it in order to target more precisely the speaker during the interview. A second microphone (AKG C1000S) can be connected as another sound source. We generally placed this transducer close to the interviewer (Figure 2). These two condensers microphones are connected to a Zoom H4n recorder which (1) delivers power to the microphones, (2) adjusts the recording level, (3) records the soundtracks in standalone, and (4) delivers good sound signals to the video camera. Indeed, we connect the Zoom H4n output to the external sound input of the video camera. It is important to control finally the sound quality by using a headphone connected to the video camera. Thus, the operator can orient the shotgun microphone and adjust the recording levels. At the end, we can obtain a film where a good soundtrack is directly synchronized with the images without an important post-processing. A copy of the soundtracks is also available on the Zoom H4n recorder.

Figure 2: Video installation for documenting languages (Karatu-Mbulumbulu-Lositete, Tanzania, Maasai Speakers, 2023)

3.3 Physiological data

The most original aspect of our project is the recording of physiological data synchronized to speech signals. These data are necessary to understand how speech sounds are produced and what are the best gestures or features necessary to describe these sounds (Demolin, 2011). A second important aspect is quantification of data which allows researchers to address the fundamental issue of variation of the studied phenomena. In the case of physiological data, we can access the roots of the variations, and not only the surface variations of the speech sound. The use of physiological data is sometimes essential to describe precisely speech sounds.

For example, the study of phonatory characteristics is facilitated by the use of electroglottography, which allows selective and direct observations, unlike the study of the speech signal, which is the result of very complex phonatory and articulatory convolutions (Figure 3). Likewise, the acoustic study of nasality is always a challenge while the physiological

mechanism of velum opening/closing remains a rather simple operation if we have a means to observe it. Measuring the nasal airflow can be a well-adapted solution for that (Figure 4). Finally, the mode of producing some obstruent consonants as plosives (egressive) or implosives (ingressive) is not so easy to detect in the speech signal. The measure of oral airflow or intraoral pressure during speech production can give quantified data about this phenomenon (Figure 5). Intraoral pressure in speech is an important physiological parameter because it highlights some complex mechanisms.

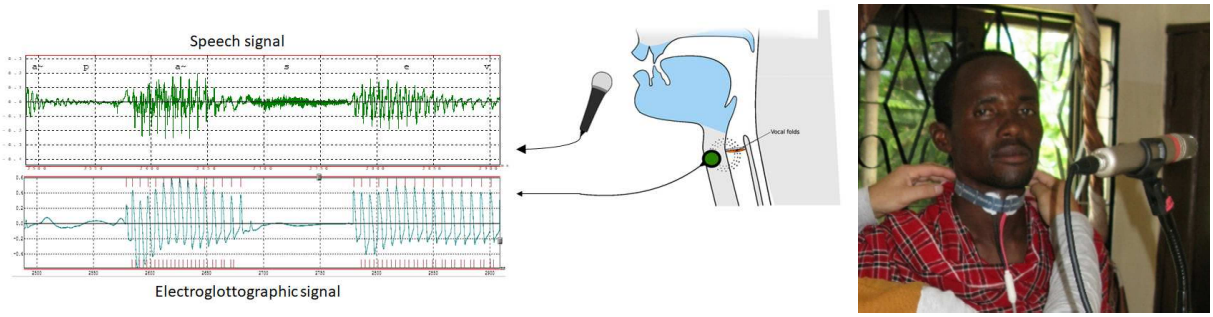


Figure 3: Selective observation of phonatory mechanisms through the electroglottography (Mto Wa Mbu, Tanzania, Maasai Speaker, 2023)

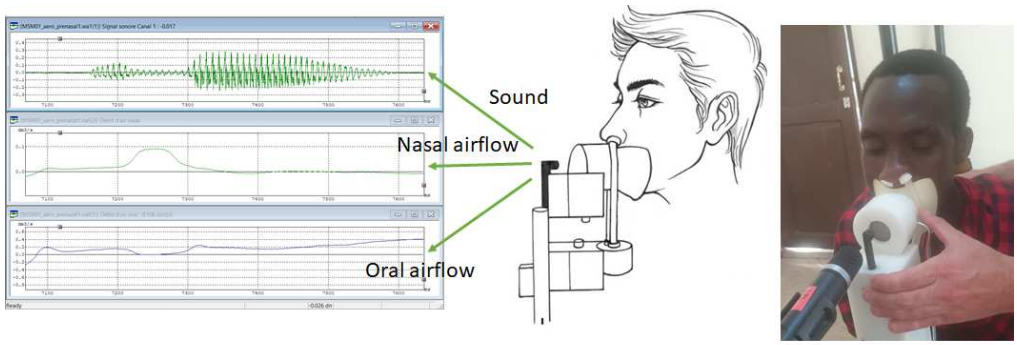


Figure 4: Measuring nasal airflow for selective observation of nasality mechanisms. Speech signal, nasal airflow and oral airflow for the Maasai word “Embaoi” (‘timber’) (Mto Wa Mbu, Tanzania, Maasai Speaker, 2023)

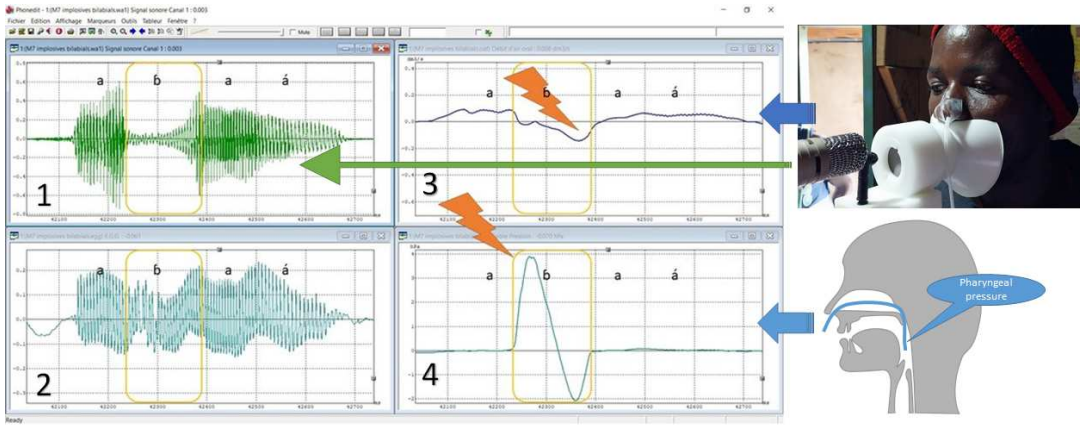


Figure 5: Multiparametric data for selective observation of implosives consonants with a Maasai speaker producing the word “aɓaá” (“to crack”). Speech signal (1), EGG (2), oral airflow (3), pharyngeal pressure (4) (Arusha, Tanzania, Maasai Speaker, 2022-2023)

During the production of an implosive consonant /ɓ/, we can see on curve n°3 a stop of the air output through the mouth. This is the occlusion mechanism. The pressure in the oral cavity increases (curve n°4) because air continues to be pushed by the lungs. Here we observe the classic process of producing a plosive. In the case of an implosive, there is an enlargement of the oral cavity and a lowering of the larynx. As the oral cavity is closed, thanks to the lips, the pressure drops suddenly and becomes negative (curve n° 4). When the lips open again on the vowel, this depression is cancelled when air enters the oral cavity.

Intraoral pressure is also interesting for the production of other consonants as, for example, ejectives in Iraqw (Demolin, Ghio et al., 2021). Getting pharyngeal pressure is helpful when we want to measure intraoral pressure regardless of the place of articulation during an oral occlusion. This involves inserting a catheter into the nasal cavity until reaching the cavum (Figure 5 bottom right). Without a medical doctor, we cannot perform this procedure ourselves. To respect this ethical constraint, we ask participants to do the insertion themselves. This operation is delicate and many participants refuse it, especially women. We always explain clearly why and how to insert the tube into the nose and before they do it we demonstrate how to do it ourselves. If speaker are not ready to do it we respect their decision. Most of the male speakers were ready to try the tube even though the insertion process sometimes failed. There is a significantly less invasive alternative for measuring intraoral pressure via the "airway interrupted method" proposed by Smitheran and Hixon (1981). The principle is to place a pressure probe in the oral cavity passing through the lips. The tube must be short and has to stop just behind the lips in order not to disturb articulation. The pressure can be measured during labial occlusions, but not for other occlusions articulated more posteriorly, which is a compromise in the field. More generally, this type of multiparametric observation help us to understand the mechanisms of speech production beyond the simple acoustic signal. Technically, we collect these data with the EVA2 workstation, developed by LPL-SQLab, Aix-en-Provence, France (Ghio et al., 2004). This equipment allows synchronized measurements of aerodynamic, electroglotto-graphic (EGG) and acoustic data. Two airflow channels and two pressure channels are available to measure oral, nasal airflow and oral pressure if necessary. The EGG signal is provided by an EG2-PCX2 model from Glottal Entreprise.

3.4 Power supply in the field

One of the major problems in using sophisticated equipment for fieldwork is the sometimes non-existent or intermittent electricity supply. This can become a real problem if power cut is frequent and it causes recording devices to malfunction (Figure 6). The solution we adopted is to work permanently with a 12V battery connected to a VICTRON inverter Phoenix 12|500 which provides a stable pure sinus electrical power supply during the recording session. We emphasize the need to use a pure sine inverter (and not a quasi-sinus inverter) because a poorly stabilized power supply can pollute the recorded signals by adding noise pulses linked to electric current chopping (Figure 7).

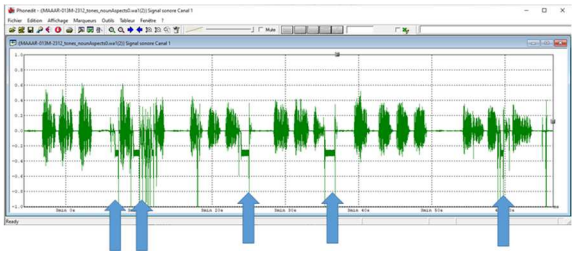


Figure 6: Effects of power supply microcuts on the sound wave (Arusha, Tanzania, 2023)

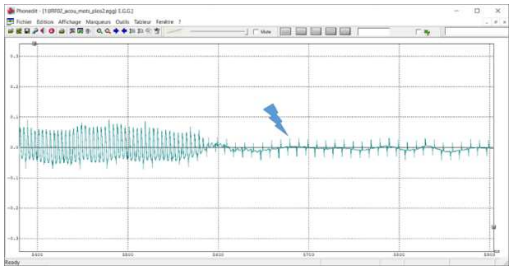


Figure 7 : Effects of a non-pure sinus inverter on the EGG signal (Kwermusl, Tanzania, 2020)

The next important question is to have a solution for charging the batteries. We have sometimes adopted the solution of portable solar panels connected to a charge controller. It is a solution that makes us completely autonomous, which is interesting in isolated areas. The disadvantage is the weight and bulk of the panels when travelling. The other solution is to have several batteries, to charge them at night or during breaks and then use them during recording sessions. Please note that it is forbidden to transport some batteries on passenger aircraft, therefore, it is necessary to find a solution with local providers.

4. Some sound systems of African Rift Valley languages

4.1 Why focus attention on the African Rift?

The Great African Rift includes several linguistic areas belonging to the four major language families of the continent: Afro-Asiatic, Niger-Kongo, Nilo-Saharan and Khoesan (Kießling et al., 2007). The linguistic diversity of this area is the result of migrations and contacts, sometimes very ancient, between populations of hunter-gatherers, pastoralists and farmers. The comparison between languages suggests several migratory phases and contacts that have repeatedly modified the linguistic landscape. The wide geographic range of some language families, such as Cushitic, which extends from Ethiopia to Kenya and Tanzania, is indicative of these ancient population movements. Some languages like Sandawe and Hadza have similar sounds, clicks, to those of the Khoesan family found in Botswana and Namibia.

The sound systems of the Rift languages have particular sound types, some of which are not

common in the world's languages. The case of non-pulmonic consonants such as ejectives, implosives, and clicks is particularly notable. Are these consonants very ancient remnants of elements that are reflected in present-day languages? Or are they the product of mechanisms of innovation and the complexification of sounds and sound systems? What is the link between the non-pulmonic consonants and the physiological mechanisms of swallowing or coughing? Is there a link between clicks and ejective consonants?

The answers to these questions require an accurate observation and description of the phoneme production mechanisms of these languages. In our project, the focus is on several languages from different linguistic families in the area (Figure 8).

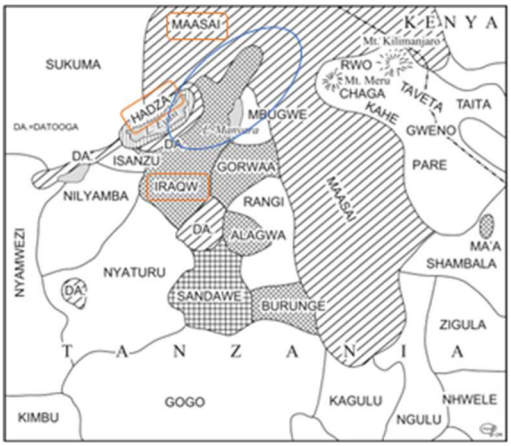


Figure 8: The languages of the Tanzanian Rift Valley Area (Source: Kießling et al., 2007)

4.2 Hadza

Hadza is a language spoken along the shores of Lake Eyasi in Tanzania by around 1,000 people (Figure 8). Traditionally, Hadzabe are full-time hunter-gatherers. Nowadays, most of them are bilingual in Swahili but Hadza language transmission to younger generation is still robust. However, there is no established Standard orthography (http://glottopedia.org).

Greenberg classified Hadza as Khoisan due to its use of click consonants but the Hadza is now considered as an isolate. The debate is delicate because if Hadza is linked to the Khoisan family, we can consider it as a remainder of the extended settlements of the Khoisan people in central Africa. If, on the other hand, as other scholars think, the Khoisan affiliation of the Hadza is not proven, we would admit that other linguistic families had existed in southern Africa, probably in contact with Khoisan, and that they finally disappeared, leaving behind only Hadza as a witness (Philipson, 2017). The precise description of the Hadza Sound System is therefore an important issue.

The consonantal system of Hadza is one of the most complex in the world, with around 60 consonants (Table 1). The inventory is controversial (Miller, 2008; Sands, 2013). It is sometimes difficult to know whether it is actually a distinct phoneme or an allophonic variation of another one.

We conducted a field mission in 2020 based in Mwangeza, a town located in the south of Lake Eyasi (https://mapcarta.com/34359060). This town is not inside the Hadza area and our speakers had to travel to this town for data recording. We recorded 4 female and 4 male speakers. Our corpus includes data on clicks, ejectives and prenasals features. We have also a spontaneous speech for a single speaker telling a story about hunting. The results on clicks are available in Demolin, Harvey et al., (2021).

Consonants		Labial	Dental	Alveolar	Lateral	Post-alveolar	Velar	Labialized velar	Radical
Clicks	Oral clicks		*c ch* /ǀ ǀʰ/	*q qh* /ǃ ǃʰ/	*x xh* /ǁ ǁʰ/				
	Nasal clicks	*(mcw)* /ʘʷ/	*nc* /ǀ̃/	*nq* /ǃ̃/	*nx* /ǁ̃/				
	Glottalized nasal clicks		*cc* /ǀˀ/	*qq* /ǃˀ/	*xx* /ǁˀ/				
Stops	Oral stops	*b p (ph)* /b p pʰ/		*d t (th)* /d t tʰ/			*g k (kh)* /g k kʰ/	*gw kw (khw)* /gʷ kʷ kʷʰ/	*' (ʔ)* /ʔ/
	Prenasal stops	*(mb mp)* /mb mpʰ/		*(nd nt)* /nd ntʰ/			*(ng nk)* /ŋg ŋkʰ/	*(ngw)* /ŋgʷ/	
	Nasal stops	*m* /m/		*n* /n/		*(ny)* /ɲ/	*(ng')* /ŋ/	*(ng'w)* /ŋʷ/	
	Ejectives	*(bb)* /pˀ/		*zz* /ʦˀ/	*dl* /ɗɮˀ/	*jj* /ʧˀ/	*gg* /kxˀ/	*ggw* /kxʷˀ/	
Affricates	Oral affricates			*z ts tsh* /ʣ ʦ ʦʰ/	*tl tlh* /ɗɮ ɗɮʰ/	*j tc tch* /ʤ ʧ ʧʰ/			
	Prenasal affricates			*(nz ns)* /nʣ nʦʰ/		*(nj)* /nʤ/			
Fricatives	Fricatives	*f* /ɸ/		*s* /s/	*sl* /ɬ/	*sh* /ʃ/			*(hh)* /ʜ/
Approximants	Sonorants				*l* /ɾ/	*y* /j/		*w* /w/	*h* /ɦ/

Table 1: Hadza consonants (adapted from Miller, 2008)

4.3 Iraqw

Iraqw is a South Cushitic language spoken in the Manyara and Arusha regions of Tanzania by more than half a million speakers (Kießling et al., 2007). Iraqw people are mainly agriculturalists. Iraqw is one of the southernmost Cushitic languages in the Afro-Asian phylum (Figure 8).

The phonetic nature of the Iraqw sound system (Table 2) is (1) a series of ejective consonants, (2) a long set of unvoiced fricative consonants, (3) a contrast between modal and labialized consonants.

	Labial	Alveolar	Lateral	Palatal	Velar	Uvular	Pharyngeal	Glottal
Voiced Plosive	b	d		ɟ <j>	g gʷ			
Voiceless Plosive	p	t		c <ch>	k kʷ			ʔ ⟨'⟩
Ejective Stop						qʼ qʷʼ		
Ejective Affricate		tsʼ	tɬʼ ⟨tlʼ⟩					
Voiced Fricative							ʕ ⟨/⟩	
Voiceless Fricativ	f	s	ɬ ⟨hl⟩	ʃ ç ⟨sh⟩	x xʷ		ħ ⟨hh⟩	h
Nasal	m	n		ɲ ⟨ny⟩	ŋ ŋʷ			
Liquids		r	l					
Approximants	w			j ⟨y⟩				

Table 2: Iraqw consonants (adapted from Mous, 1993)

The ejective status of /qʼ/ is under debate (Demolin, Ghio et al., 2021). This question relating to ejectives is interesting because ejectives are also in the Hadza sound system (Table 1) as well as among the Sandawe, another click language spoken in the south of the area (Figure 8). Knowing that / tɬʼ/ is only found in 1% of the world languages (Phoible, 2019), a presence in 3 geographically close languages is probably due to a contact effect which is yet to be studied.

We conducted a first field mission in 2020 based in Kwermusl, a town located in the south of Mbulu district (https://mapcarta.com/N5033497899). This village is the heartland of the Iraqw area and we recorded 5 female and 5 male speakers. Our corpus includes data on ejectives, pharyngeal, glottal articulation, labialized vs modal contrast, fricatives, and vowels. We also had a single speaker reading the translation of the story in Iraqw "the wind and the sun". We added two male speakers in Arusha in June 2022. We completed the corpus with three male and three female speakers recorded in Mto Wa Mbu (https://mapcarta.com/34353418) at the border between the Iraqw area and Maasai land in December 2022. We have some video data to study the difference between labial movements involved in the production of consonants /ŋʷ, kʷ, gʷ, qʷˊ, xʷ/ and compare them with the gestures of the bilabial nasal [m] and the labiovelar approximant [w] (Ghio et al., 2021).

4.4 Maasai

Maa is a Nilotic language spoken in Southern Kenya and Northern Tanzania by 1.5 million Maasai people. Maasai steppe covers a large area in both Kenya and Tanzania and several dialects can be distinguished: Samburu, Ilchamus, Ilkeekoŋokie, Purko, Ilwasiŋgiʃu, Arusa, Kisongo, Parakuyo. Traditionaly, Maasai are pastoralists but some sections are also farmers, especially Arusa and recently during the fieldwork in December 2023, we noticed that Kisongo too in Monduli District have started subsistence farming. In its phonological system (Table 3), Maa has a complete set of implosives, a complete system of vowels [+/- ATR], tones, and a fortis/lenis contrast for glides /j/ vs /jː/, /w/ vs /wː/, and also for rhotic /ɾ/ vs /r/. There is also a very elaborate set of interjections, ideophones and animal calls that include sounds not described in the International Phonetic Alphabet (Andrason et al., 2023; Andrason et al., 2021; Karani et al., 2022). The implosive mechanism is not easy to observe and the use of aerophonometry, described in §3.3, is necessary.

		bilabial	alveolar	postalveolar	palatal	labiovelar	velar	glottal
plosive	explosive	[p] [b]	[t] [d]				[k] [g]	ʔ
	implosive	[ɓ]	[ɗ]				[ɠ]	
affricate	explosive			[t͡ʃ] [d͡ʒ]				
	implosive			[ʃ]				
fricative			[s]	[ʃ]				[h]
nasal		[m]	[n]		[ɲ]		[ŋ]	
rhotic	tap		[ɾ]					
	trill		[r]					
lateral			[l]					
approximant /glide	lenis				[j]	[w]		
	fortis				[jː]	[wː]		

Table 3: Maasai consonants (Karani et al., 2023)

We conducted 3 field missions in 2022-2023. The first location was Ilkurot village around Arusha (https://mapcarta.com/N10836698430) where the Arusa dialect is spoken. We recorded 14 male and 8 female speakers for phonetic purposes (implosives, vowels, approximants, tones, ideophones…). Moreover, we video-recorded twice additional speakers telling stories. The second location was Esilalei and Selela villages where Kisongo dialect is spoken (https://mapcarta.com/N946585023). We recorded 8 male and 4 female speakers for phonetic purposes. Likewise, we video-recorded additional speakers telling stories in the villages. In order to study the contact between Maasai and Iraqw people, we conducted a mission in Lositete (https://mapcarta.com/N7512233367) where we recorded video clips of 5 speakers telling stories with the specificity of being Arusa Maasai surrounded by Kisongo Maasai and Iraqw. Some preliminary results on implosives have been shared by Demolin et al. (2022).

5. Conclusion

The necessity of empirical data to quantify observed phenomena is part of the scientific approach. This is the case in linguistics particularly in African Linguistics. The examples given before show that phoneticians have to obtain sets of quantitative data in order to understand patterns of variation. The Rift Valley in Tanzania is a perfect area to observe various and original linguistic features. With the experience we gain in the missions doing fieldwork attests that 'dirty feet' fieldwork is not a walk in the park but certainly feasible and it can be very successful if researcher prepare well. Since some linguistic data are rare and sometimes difficult to collect, data sharing is an important aspect that linguists need to take seriously during data collection and after fieldwork. Our fieldwork missions have been successful to a large extent because one of the collaborators is a trained linguist who is a native speaker of one of the languages under investigation. Hence, we recommend working with local researchers to maximize the chances of succeeding in fieldwork missions.

6. Acknowledgments

This study was supported by grant CNRS “SYSORI” (Dispositif de Soutien aux Collaborations avec l’Afrique subsaharienne , 2021) and by grant CNRS “COSYSORI” (Africa Visiting Fellowships Programme, 2023). We thank the CEP (https://cep.lpl-aix.fr/) for the loan of equipment used in the field. We thank Andrew Harvey and Richard Griscom for their participation in the recordings with the Hadza. We thank Maarten Mous for his participation in the recordings with Iraqw. We thank L Jalby, from OZO company (https://ozo-electric.com/en/) for their advice and products relating to the supply of electricity using solar equipment and batteries.

7. Bibliographical References

Adda, G., Stüker, S., Adda-Decker, M., Ambouroue, O., Besacier, L., Blachon, D., Bonneau-Maynard, H., Godard, P., Hamlaoui, F., Idiatov, D., Kouarata, G.-N., Lamel, L., Makasso, E.-M., Rialland, A., Van De Velde, M., Yvon, F., & Zerbian, S. (2016). Breaking the Unwritten Language Barrier : The BULB Project. *Procedia Computer Science*, 81, 8-14. https://doi.org/10.1016/j.procs.2016.04.023

Andrason, A. & Karani, M. (2023). Emotive interjections in Arusa Maasai. Italian Journal of Linguistics, 35(2), 75-118. https://doi.org/10.26346/1120-2726-212

Andrason, A. and Karani, M. (2021). Conative calls to animals: From Arusa Maasai to a cross-linguistic prototype. *Łódź Papers in Pragmatics*, 17(1), 3–40. https://doi.org/10.1515/lpp-2021-0002

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., & Rialland, A. (2016). Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App. *Procedia Computer Science*, 81, 61-66. https://doi.org/10.1016/j.procs.2016.04.030

Boyd, M., Mazaudon, M. , Michaud, A., Guillaume, S., François, A, Adamou, E (2014). Documenting and Researching Endangered Languages: The Pangloss Collection. *Language Documentation & Conservation*. 8: 119-135.

Crowley, T. (2007), Field Linguistics. A Beginner’s Guide, Oxford, Oxford University Press.

Demolin, D. (2011). Aerodynamic techniques for phonetic fieldwork. *ICPhS International Congress of Phonetic Sciences*, Aug 2011, Hong Kong, China. ⟨hal-00646103⟩

Demolin, D., Ghio, A., Mous, M. (2021). A phonetic study of Iraqw ejectives consonants. *10th World Congress of African Linguistics (WOCAL)*, Leiden, Netherlands. ⟨hal-03156237⟩

Demolin, D., Harvey, A., Griscom, R., & Ghio, A. (2021). Acoustic features of Hadza clicks. *The Journal of the Acoustical Society of America*, *150*(4_Supplement), A68-A68. https://doi.org/10.1121/10.0007649

Demolin, D., Ghio, A., Karani, M. (2022). Acoustic, aerodynamic and articulatory features of Maasai implosives. *Colloquium on African Languages and linguistics*, Leiden, Netherlands ⟨hal-03767661⟩

Gasquet-Cyrus, M. (2015). I come and go between fieldwork: Thoughts on Fieldwork in Sociolinguistic Theorization. *Langage et société*, 154, 17-32. https://doi.org/10.3917/ls.154.0017

Ghio, A., & Teston, B. (2004). Evaluation of the acoustic and aerodynamic constraints of a pneumotachograph for speech and voice studies. *International Conference on Voice Physiology and Biomechanics*, Univ. Méditerranée, 2004, Marseille, France. pp.55-58. ⟨hal-00142982⟩

Ghio, A., Mous, M., & Demolin, D. (2021). Labialized consonants in Iraqw. 10th World Congress of African Linguistics. ⟨hal-03156276⟩

Güldemann, T. (Éd.). (2018). The Languages and Linguistics of Africa. De Gruyter. https://doi.org/10.1515/9783110421668

Karani, M., & Andrason, A. (2022). Ideophones in Arusa Maasai : Syntax, morphology, and phonetics. *Open Linguistics*, *8*(1), 440-458. https://doi.org/10.1515/opli-2022-0220

Karani M., Andrason A. (2023), A Living Grammar Sketch of Arusa, *Living Tongues Institute for Endangered Languages* https://doi.org/10.5281/ZENODO.10389820

Kießling, R., Mous, M., & Nurse, D. (2007). The Tanzanian Rift Valley area. In B. Heine & D. Nurse (Éds.), A Linguistic Geography of Africa (p. 186-227). Cambridge University Press. https://doi.org/10.1017/CBO9780511486272.007

Maddieson, I. (2002). Phonetics in the Field. *Annual Meeting of the Berkeley Linguistics Society*, 28(1), 411. https://doi.org/10.3765/bls.v28i1.3855

Miller, K. (2008). Hadza grammar notes. *Symposium on Khoisan Languages and Linguistic: In Memory of Professor Anthony Traill*, Riezlern.

Mous, M. (1993). A Grammar of Iraqw. Helmut Buske Verlag.

Phillippson, G. (2017). Langues et histoire dans le Rift. In *Le Rift est-africain* (IRD Éditions, p. 367-376). https://books.openedition.org/irdeditions/1790

Sands, B. (2013). Phonetics and Phonology: Hadza. In R. Vössen (Ed.). The Khoesan languages. London. Routledge.

Smitheran, J. R., & Hixon, T. J. (1981). A clinical method for estimating laryngeal airway resistance during vowel production. *The Journal of Speech and Hearing Disorders*, *46*(2), 138-146. https://doi.org/10.1044/jshd.4602.138

8. Language Resource References

Atlasi ya lugha za Tanzania. (2009). Mradi wa Lugha za Tanzania, Chuo Kikuu cha Dar es Salaam. ISBN: 978-9987-691-26-5

PHOIBLE 2.0 as CLDF dataset (v2.0), Moran, S., & McCloy,(2019). https://doi.org/10.5281/ZENODO.2593234

Kallaama: A Transcribed Speech Dataset about Agriculture in the Three Most Widely Spoken Languages in Senegal

Elodie Gauthier, Aminata Ndiaye, Abdoulaye Guissé
Orange Innovation, Jokalante SARL, École Polytechnique de Thiès
Lannion, France, Dakar, Sénégal, Thiès, Sénégal
elodie.gauthier@orange.com, amina.ndiaye@jokalante.com, aguisse@ept.sn

Abstract

This work is part of the Kallaama project, whose objective is to produce and disseminate national languages corpora for speech technologies developments, in the field of agriculture. Except for Wolof, which benefits from some language data for natural language processing, national languages of Senegal are largely ignored by language technology providers. However, such technologies are keys to the protection, promotion and teaching of these languages. Kallaama focuses on the 3 main spoken languages by Senegalese people: Wolof, Pulaar and Sereer. These languages are widely spoken by the population, with around 10 million of native Senegalese speakers, not to mention those outside the country. However, they remain under-resourced in terms of machine-readable data that can be used for automatic processing and language technologies, all the more so in the agricultural sector. We release a transcribed speech dataset containing 125 hours of recordings, about agriculture, in each of the above-mentioned languages. These resources are specifically designed for Automatic Speech Recognition purpose, including traditional approaches. To build such technologies, we provide textual corpora in Wolof and Pulaar, and a pronunciation lexicon containing 49,132 entries from the Wolof dataset.

Keywords: speech dataset, Senegalese languages, low-resource setting, agriculture

1. Introduction

While information and communication technology is essential for many to thrive, 6 billion people still lack access to broadband, 4 billion lack access to the Internet, and 2 billion lack access to a mobile phone (Zelezny-Green et al., 2018). Latest estimations from UNESCO Institute for Statistics (2023) indicates around 213 million adults (population over 15 years old) who could not read or write, in 2022, across the sub-Saharan African region, including nearly 49 million young people (15-24 years old). In Senegal, ANSD (2021) reports an overall illiteracy rate of 48,2%, reaching 62,7% in rural area.

Literacy rate relates to the official language of a country. In Senegal, the official language is French but is seldom spoken by the population in their daily lives. Senegalese people primarily use their native languages or Wolof, as a vehicular language, to communicate. World Bank (2021) reports that nearly 65% of Senegalese who do not use the Internet are hindered by a lack of digital literacy. This is partly due to the limited (if not none at all) availability of content in the language they speak. Currently, there is a severe lack of accessible content for those who do not speak the official languages in Africa. The development of technologies and tools for the most widely spoken languages would enable a larger proportion of the Senegalese people to use smartphones and applications, and to access content that is still unavailable today.

Research work as Medhi et al. (2011) and the success of WhatsApp voice communication show that the development of conversational voice services in local languages is a credible and promising way of making services more accessible. Aker (2011) also suggested in that time that combining a voice-based approach with information that can be accessed through answers to common farmer questions would overcome literacy challenges due to the common texting modes. To make progress in this area, robust speech recognition systems need to be designed for these languages. While automatic speech recognition (ASR) technologies tend to be mature in the languages most commonly found on the Web, there is still very few solutions dedicated to African languages.

In Senegal, Wolof, Sereer, Pulaar, Joola, Malinké and Soninké languages are recognised as national languages in the Constitution, but none of these six languages seriously benefit from the major technological advances generated by AI. Efforts have been made to develop speech resources and technologies in Wolof (see section 4) but no resources are available for Large Vocabulary Continuous Speech Recognition (LVCSR) in Pulaar nor Sereer. Yet, Wolof, Pulaar and Sereer languages are spoken in more than two-thirds of Senegal country (Leclerc, 2023).

Agriculture is the primary source of income for 2 billion people around the world (Zelezny-Green et al., 2018). In Senegal, 55% of the population is involved in the agricultural value chain, including family farming, livestock breeding, and fishing.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 10–19

Today, digital technologies are assisting farmers in expanding their businesses by enabling them to position themselves on marketplaces, providing them with information on commodity prices, and granting them access to suitable financial services. Nonetheless, such solutions are still not appropriate for farmers, particularly given the prevalence of written communication and the use of a language they do not speak, when interacting with these interfaces.
With the intention of speech solutions development, this dataset is intended to fill the gap in this area.

Paper contribution. This paper presents the dataset created from speech data produced and annotated during the Kallaama project, as well as textual data gathered from the web, with the aim of developing voice-based solutions in local languages.

Paper outline. After an introduction, we present the project in section 2. The targeted languages are described in section 3 and existing resources in these languages are listed in section 4. Section 5 presents the collection methodology, while section 6 gives details about the dataset. Then, we present some of the challenging times we faced during the project in section 7 and we mention some of the resulted limitations in section 8. Finally, section 9 concludes and gives some perspectives about the use of Kallaama.

2. The Kallaama project

"Kallaama" means "speech" (from Latin "verbum") in Wolof.

2.1. Description

As mentioned in section 1, no resources are available to build LVCSR systems in Pulaar nor Sereer. Only a small amount exist in Wolof, but none focus on agriculture. The Kallaama project aims to fill this gap by producing several dozen hours of transcribed and annotated localized audio data, to train speech recognition systems in three of the Senegal's main national languages: Wolof, Sereer and Pulaar.

The choice of these 3 languages was guided by the number of speakers in the country. There are around 5 million native speakers of Wolof, 3.5 million native speakers of Pulaar and 1.3 million native speakers of Sereer (Leclerc, 2023), which represent three quarter of the total population. These are the 3 most widely spoken languages in Senegal, and they are also spoken cross several borders.

The data produced are natural, spontaneous utterances, with vocabulary in context, designed to develop large vocabulary speech recognition models, particularly relating to the agricultural domain. Speech recognition is the main technological barrier to be overcome to develop voice-based services for people with little or no literacy. Agriculture plays an important role in rural activities in Senegal. It is one of the pillars of the Senegalese economy, estimated to contribute 15% of GDP in 2022 as mentioned in the Annual Agricultural Survey of DAPSA (2023), and a large proportion of the population remains directly dependent on it.

2.2. Use case

Several and local companies and start-ups in the IT sector are increasingly embarking on the production of AI solutions that take national languages into account. These are essentially automatic text or speech translation solutions, allowing them to expand their customer base and offer their services in French or English to users who prefer Wolof, Pulaar, Sereer, or other languages. Serious initiatives also have been noted in the development of multilingual chatbots and voicebots. However, due to the scarcity of natural language data in local languages, most of them rely on synthetic data from machine translation systems. Moreover, AI models still only marginally address agriculture. Yet, digital solutions for agricultural extension work cover a range of needs, including information delivery services, small business management tools, training and skills enhancement, and financial services (Zelezny-Green et al., 2018).

The Kallaama dataset contributes to the growth of the agricultural sector in Senegal. It can strengthens food security by providing vital information directly in the farmers' native language, through the development of voice-based services such as personalised agricultural and financial advice to smallholder farmers. Besides, the produced transcriptions increase the available datasets in Senegalese languages, and will boost AI-based developments for agriculture, like setting up knowledge bases, conversational assistants, recommendation systems and decision support systems.

3. Focus on the targeted languages

Wolof, Pulaar and Sereer languages are spoken by nearly 80% of native speakers in Senegal. Cissé (2005) indicates 43,7% of Wolof native speakers, 23,2% of Pulaar native speakers and 14,8% of Sereer native speakers. These three languages belong to the Niger-Congo phylum and are part of the North-Atlantic family group. They are toneless,

unlike most Niger-Congo languages[1]. By having a national status, the three languages received an official spelling system. It is based on the Latin characters.

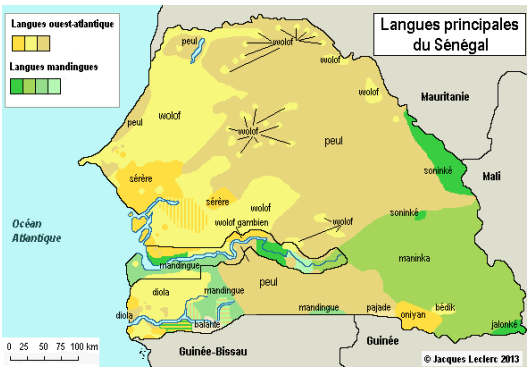


Figure 1: Map of main languages spoken in Senegal (Leclerc, 2023)

3.1. Wolof

Wolof is by far the most spoken language in Senegal. It is the native language of about 5 millon speakers. The Wolof spoken in Senegal is identified by the ISO 639-3 language code name "wol"[2]. By being spoken by almost 90% of the population, Wolof is the national language of communication, widely surpassing French in terms of usage (Cissé, 2005). On social networks, comments are mainly written in Wolof in response to articles written in French. The National Assembly provides a translation service as 20% of MPs do not speak French. Additionally, private TV and radio channels have developed programmes in Wolof (OIF, 2022).

3.2. Pulaar

Pulaar is part of the Fulfulde languages. Fulfulde is spoken in about 20 sub-Saharan African countries, by nearly 30 million people, from West to Central Africa. "Pulaar" refers to the variant spoken in Senegal. Pulaar is the native language of about 3.5 millon of the Senegalese people, making it the second most widely spoken language in the country. Pulaar speakers across the country do not always understand each other. Clear differences in accents and lexicons should be noted. There may be borrowings and mutual influences between the accents. The Pulaar spoken in the north, considered as the reference in Senegal, used in the areas of Fouta Toro and Ferlo, is different from the one spoken in the south, in Fouta Djallon and Boundou areas, and from the centre (particularly in Saloum). The economic activities practiced by the Fulani in these regions are at the origin of these differences, without forgetting the mobility of populations and inter-cultural exchanges.

3.3. Sereer

Sereer language is spoken by around 1 million speakers, making it the third language spoken in Senegal. Several dialects are spoken in Senegal (Renaudier, 2012), and mutual understanding between Sereer speakers is sometimes difficult. The majority of the recordings proposed in this dataset are in Sereer-Siin (ISO 639-3 code "srr") variant, which is spoken in a region between the Petite Côte (south of Dakar) and the Gambia, and which is considered as "standard" Sereer. Nonetheless, depending on where the recording was made, it may be in another variant. The official script is based on the standard Sereer-Siin variant but is very little used for writing. The language is fundamentally spoken.

4. Existing language resources for Wolof, Pulaar and Sereer

More material (in any field of application, from linguistic description to language learning) can be found in Wolof, as a vehicular language. The situation is very different with Pulaar and Sereer: as vernacular languages, they are mainly spoken and rarely written. The presence of Wolof online is preponderant against Pulaar and Sereer, reflecting its place in the Senegalese society.

To build voice-based solutions, very few datasets were released so far in Wolof. Pulaar and Sereer speech datasets are nearly non-existent, exceptions made from the initiatives presented below.

Before this work, the largest transcribed speech dataset in Wolof was the one collected by Gauthier et al. (2016). It consists in 18 hours of validated read sentences.

Wolof is also proposed in FLEURS, a multilingual dataset consisting in translation of English sentences that has been read by native speakers (Conneau et al., 2023).

Nelson (2022) conducted a project of large collection of Wolof speech consisting in 519 hours of audio recordings, for which 6.45 hours have been transcribed so far. Among the 2,018 Wolof transcriptions, we counted 608 translations in Pulaar, 571 in Sereer but only 156 audio recordings translated in both languages.

Finally, the last significant work we found involving the three Senegalese languages addressed in this

[1]As mentioned by Creissels (2019), non tonal languages are primarily spoken in the Atlantic languages of western Senegal and the Bantu languages of eastern Kenya and Tanzania (like Swahili).

[2]Another variant of Wolof is spoken in Gambia, for which the ISO code is "wof".

paper, is part of a data collection project of isolated words for keyword spotting, led by the Senegalese Galsen AI community (Djiba, 2021).

5. Collection methodology

5.1. Audio recordings and transcriptions

Audio recordings. The recordings are about agriculture. The recorded consist of farmers, agricultural advisers, and agri-food business managers. All the data is produced by Jokalante, a Senegalese company specialising in the dissemination of information about agriculture in local languages. Type of recordings comprise interactive radio programmes, focus groups, voice messages, push messages and interviews[3]. Therefore, spontaneous speech is prevailing. Quality of audio may vary depending on the type of programme. For instance, focus group are made outdoor and so noises may arise from the outside (cars, wind, birds, additional voices). In radio programmes, music and jingles sometimes also appear. A selection from these recordings were transcribed, resulting in over a hundred hours of spontaneous speech in the three targeted languages (see Section 6.1).

Transcriptions. To produce written form of the audio recordings, we asked the transcribers to follow the rules edited by the Centre of Applied Linguistics of Dakar (CLAD)[4], which coordinates the orthographic standardization of the national languages in Senegal. Despite that caution, it was very difficult to obtain a standard form in the writing of languages concerned in the present work. As mentioned by Robert (2022) for Wolof, official rules are rarely used by the population (as example, advertisements are often written with alternative forms), even if an official orthography is established since 1971. The same situation appears for Pulaar and Sereer, and this is primarily due to the fact that the national languages are taught very little in the education system. In addition, the transcription work involved recordings of spontaneous speech, making the work all the more complex and time-consuming. Transcription task was performed by 3 students in Linguistics, in the language they natively speak. They used the dedicated Transcriber[5] tool to achieve the task. The work took 9 months to complete. Then, 3 qualified experts, specialised in the languages of the transcripts, reviewed a subpart of the transcriptions produced by the students. At first, we were aiming to verify half of the transcriptions produced, for each language. But it was an ambitious goal given the complexity and arduousness of the task required. Nonetheless, nearly 13 hours of speech transcription were checked in Wolof, 11 hours in Pulaar and 11 hours in Sereer within the allotted time.

5.2. Texts collection

Senegalese languages are low-resourced. Very few data in the targeted languages were unearth. First of all, no documents on agriculture were found. Since the observations from Renaudier (2012) who mentioned that the local press is predominantly written in French, with only a few newspapers available in Wolof, and that no press were available in Sereer at the time, the situation remains unchanged.

Most of the written sources found were in Wolof. The Wolof corpus we distribute is composed of the books, Wikipedia articles (dump from summer 2023), the first book of the New Testament, two historic blogs about Hubert Fichte (a German novelist) and Cheikh Ahmadou Bamba (a theologian), publicly available online articles from newspapers. Open source data from the Programme Algorithme et Solution (PAS) challenge[6], organised by the Institution des Algorithmes du Sénégal (IAS), have also been included. Written data in Wolof can also be retrieve from open source research projects (in particular, the ALFFA project[7] and Masakhane[8]). We choose not to add them to our release as they are already clean and easy to get.

We found a very little amount of writings in Pulaar[9]. We extended our data research to include varieties spoken in regions bordering Senegal, and finally found more websites written in this language, particularly in Mauritania.

About Sereer, although it is the third most widely spoken language in Senegal, gathering written data poses a significant challenge. Despite extensive research, no textual content was found on the consulted websites. We even went to the two main university of linguistic and language libraries in Dakar (Cheikh Anta Diop University (UCAD) and to the Institut Fondamental d'Afrique Noire (IFAN)), and only found two books written in Sereer. We still tried to apply some Optical Characters Recognition (OCR) tools to convert it into digital texts, but the special characters existing in Sereer were not recognised.

We have deliberately excluded all social networks in order to avoid biases that could be

[3] For each dataset, the number of recordings per programme type is detailed in appendix C.

[4] http://clad.ucad.sn/

[5] http://trans.sourceforge.net/

[6] https://www.ias.sn/pas/

[7] https://github.com/getalp/ALFFA_PUBLIC/

[8] https://github.com/masakhane-io/masakhane-ner/

[9] Without distinction of dialectal variants spoken in Senegal.

induced in future models. For example, Dione (2016) observed, in her study on the online usage of Wolof and Sereer languages, that most internet users use French and Wolof alternatively in a single message. Besides, Wolof and Pulaar are the only two national languages present on the websites consulted by the author, while Sereer is also his subject of study. Moreover, the author indicates that internet users use Wolof to criticise, to display political choices and connivance, and to insult. For all these reasons, we preferred not to collect textual data from the forums.

These text corpora can be used to train monolingual and multilingual language models on the theme of agriculture. Language models are involved in various NLP tasks, such as ASR rescoring or natural language understanding/generation (NLU/NLG) modelling.

5.3. Lexicon

We found no dictionary with word pronunciation for Pulaar and Sereer, so we could not train a grapheme-to-phoneme (G2P) model for these languages. For Wolof, we used the lexicon from ALFFA project to train a G2P model, in order to generate phonetic transcription of the Wolof speech set. The G2P model was trained using Phonetisaurus[10]. The generated phonetic symbols are in X-SAMPA alphabet. We provide the G2P model and the lexicon in a GitHub repository. It can be useful to train HMM-based ASR models.

6. Dataset details

The dataset is released under the CC-BY 4.0 license. All textual data (transcriptions, text corpus, lexicon) are available on GitHub[11]. Audio recordings are hosted on both OpenSLR[12] and Zenodo[13] platforms.

6.1. Audio recordings and transcriptions

Audio files have been converted into 16 kHz, 16-bit, mono channel, to fit the standard format used in ASR. Transcriptions are provided under the original Transcriber format (.trs), as well as in stm NIST format (.stm) as this one is more often used by ASR toolkits.
Details about speech datasets are given in table 1.

Language set	Total Duration	#Turn-taking	Gender (%)	
			F	M
Wolof	55h12	46,907	10.2	89.8
Pulaar	31h55	16,558	13.6	86.4
Sereer	38h12	9,007	28.0	72.0
Overall	125h19	72,472	17.3	82.7

Table 1: Kallaama speech corpus overview

The high number of turn-taking that can be observed in the table 1 for the Wolof set is explained by a larger amount of interviews and focus group, involving more people in the talk.
The underrepresentation of women's voices in this corpus is regrettable, but it reflects the interviews conducted and the women's presence in agribusiness.
More details are given in appendix B, where we also describe the checked subpart of the dataset.

6.2. Texts collection

The set of texts collected in Wolof, before the application of post-processing methods, totalled 3,244,642 words. The set of texts collected in Pulaar, before the application of post-processing methods, totalled 5,462,823 words. As we said in subsection 5.2, no written data were found in Sereer.
During post-processing non roman characters were removed from raw texts. Punctuation has been preserved to give users greater freedom, depending on how the corpus will be used. Finally, a new line was added after each final punctuation mark (the dot, exclamation and question marks) while spaces was added between other kind of typography mark (such as comma, colon, semi-colon, dash, bracket, etc.).
After these post-processing steps, the Wolof text corpus contains 1,140,508 words, while the Pulaar text corpus contains 742,024 words. Detailed are given in table 2 and table 3, for Wolof and Pulaar respectively. This considerable reduction in content reflects the significant presence of other languages in the writings, particularly French and Arabic (we did not apply a language identification algorithm, but we did remove many characters in the Arabic alphabet). We also found a quite large number of Cyrillic characters in the collected texts from Wikipedia.
The compiled data will enhance the understanding of the usage of the languages and strengthen the ability to develop more robust linguistic tools. It will also serve as a training baseline for language models.

[10] https://github.com/AdolfVonKleist/Phonetisaurus/
[11] https://github.com/gauthelo/kallaama-speech-dataset/
[12] https://www.openslr.org/151/
[13] https://zenodo.org/records/10892569/

Sources	#Words	Distribution
Newspapers	571,122	50%
Wikipedia	346,604	30%
PAS Challenge	157,119	14%
Book	27,283	2%
New Testament	22,468	2%
Blog	15,912	1%
Overall	**1,140,508**	**100%**

Table 2: Details about the web scrapped texts in Wolof, after cleaning

Sources	#Words	Distribution
Newspapers	698,400	94%
Blog	43,624	6%
Overall	**742,024**	**100%**

Table 3: Details about the web scrapped texts in Pulaar, after cleaning

6.3. Lexicon

In the aim to build ASR systems, we also provide a pronunciation dictionary for Wolof. It contains 49,132 phonetised entries from speech transcriptions and texts. Entries can also be loanwords, such as French words, since code-switching is frequent in Senegal and therefore occurs in the speech dataset. Entries are phonetically transcribed with the X-SAMPA characters.

7. Challenges encountered

Transcribers struggled to write some of the words because of the absence of certain characters on standard keyboards, such as ɓ (b-hook), ƈ (c-hook), ɗ (d-hook), ƥ (p-hook), ƭ (t-hook) which exist in the spelling of African languages, in particular in Pulaar and Sereer. The SenLangEdit visual keyboard application[14], especially developed to write the national languages of Senegal, still eased the transcription process.

7.1. Writing rules

It was particularly hard to find qualified experts for checking the quality of the produced transcriptions. We ask each expert to make a report of their reviews. For Wolof, the expert declared that the work was quite easy. To complete the work within the allocated time, he managed to check nearly 13 hours of speech transcriptions and the conclusion was very encouraging. He noted a very good quality of work, with very few mistakes. In contrast, the two qualified experts hired to review the transcriptions in Pulaar and Sereer declared a tedious work. In spite of this, they manage to verify around 11 hours of audio recordings each. They raised numerous mistakes and warned us that their work would be more about rewriting than simple checking and correction. We detail the main mistakes found in appendix A.

In fact, this assessment of the transcriptions quality primarily indicates a lack of written skills rather than a lack of attention to transcription quality. This is the result of attempting to transcribe a language that has traditionally been unwritten. As long as these languages are not taught to be write, there will be no good written productions.

7.2. Spoken dialects

The fact that recordings are produced throughout the Senegal[15] made the transcription work much more complex, because of the several dialectal variations of Pulaar and Sereer which are spoken in the country (see sections 3.2 and 3.3). We selected the audio recordings at the very beginning of the project, before the transcriber hiring. But Pulaar or Sereer transcribers sometimes listened to programmes recorded that they did not understand because of a conversation in a dialect that they do not speak. Consequently, we had to carry out a second recordings collection campaign that took into account the specific dialects spoken by the transcribers.
This process also highlights the need to take account of the semantic subtleties between dialectal varieties, especially when dealing with a particular subject (in this case agriculture, but it could be health or finance) and illustrates the challenges inherent in accurately and exhaustively preserving the meaning of words in these languages.

8. Limitations

Transcription work is a very challenging task, and to produce a transcript from spontaneous speech, when overlapping events occur, sometimes in noisy environments, is even more so. Add to this the use of specialised software that is unfamiliar to the workers, with keyboards that are not adapted to writing the language, and the start of the work becomes even more tedious.
Despite the care of all the workers involved in this project, this dataset contains some transcription mistakes, and the spelling used may not correspond exactly to the expected standards, as pointed out in Section 7. Only one transcriber was selected

[14] https://esp.sn/senlangedit-un-clavier-virtuel-pour-la-promotion-des-langues-nationales/

[15] Diouroup, Fatick, Kaffrine, Kaolack, Kebemer, Kelle, Koungheul, Louga, Matam, Ndoundour, Niodior, Nioro, Podor, Saint-Louis, Tambacound.

per language to carry out the transcription work. Perhaps some mistakes could have been avoided if more transcribers were doing the job (supposing this is possible, since the number of skilled people is very limited). But, due to production costs, we have chosen to provide the community with a larger set of transcribed data rather than increasing the number of transcribers. In this way, we have been able to increase the number of subjects covered on agriculture. A larger speech dataset is also more suitable for large-scale studies, such as phonetic and phonological research on Atlantic languages, a field where works lack.

9. Conclusion and Opportunities

In this paper, we present the work carried out to create a transcribed speech dataset on Wolof, Pulaar and Sereer, the 3 most widely spoken languages in Senegal. This dataset comprises 55h of audio recordings in Wolof, 32h in Pulaar and 38h in Sereer, all along with their corresponding transcriptions. We also provide more generic text corpora in Wolof and Pulaar, as well as a Wolof phonetic lexicon along with its G2P model. These resources can be used for setting up traditional ASR systems.
As pointed out by many recent studies (Joshi et al., 2020; van Esch et al., 2022; Ruder et al., 2022; Adebara and Abdul-Mageed, 2022), a lot of languages with large speaker populations still are under-represented in natural language processing (NLP) studies and applications, reinforcing inequalities such as knowledge access. We hope that this work will stimulate interest in the development of applications that incorporate the vernacular languages of Senegal, but also that it will be a source of inspiration and encouragement to develop the same kind of resources in order to progress towards the inclusion of languages in the world of AI.

Opportunities offered by this dataset are numerous. From a scientific perspective, the speech dataset released can be exploited for instance, to study phonetic phenomena occurring in a spontaneous context, to study speech interaction, or to study the impact of spontaneous and noisy speech on recognition systems. From a technical perspective, this dataset can be used to solve various AI tasks, including speech modelling (such as speech-to-text or spoken language understanding), automatic response modelling (as QA answering), and language modelling (used from scratch or used to fine-tuned a pre-trained multilingual model). From a technological perspective, it can be utilised to develop speech recognition systems, generic or specific to the agricultural sector, as well as localised conversational agents for answering questions on agricultural topics related to the Senegal context and in national languages.

10. Acknowledgements

The Kallaama project was made possible thanks to the financial support from the Lacuna Fund[16], the world's first collaborative effort to fund labelled data for social impact. Lacuna Fund promotes creation, expansion, and maintenance of labelled datasets in three domain areas with key needs: agriculture, health, and languages.
The authors would also like to acknowledge and thank the linguists and data scientist interns involved in the project.

11. Bibliographical References

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Jenny C Aker. 2011. Dial "A" for agriculture: a review of information and communication technologies for agricultural extension in developing countries. *Agricultural economics*, 42(6):631–647.

ANSD. 2021. Enquête harmonisée sur les conditions de vie des ménages (EHCVM). Technical report, Agence Nationale de la Statistique et de la Démographie (ANSD).

Mamadou Cissé. 2005. Langues, état et société au Sénégal. *SudLangues. Revue électronique internationale de Sciences du langage*, 5(1):99–133.

Denis Creissels. 2019. Morphology in Niger-Congo languages. In *Oxford Research Encyclopedia of Linguistics*.

DAPSA. 2023. Rapport de l'Enquête Agricole Annuelle (EAA) 2022-2023. Technical report, Direction de l'Analyse, de la Prévision et des Statistiques Agricoles.

Amadou Dione. 2016. *Contacts et valorisation du sérère et du wolof, langues nationales du Sénégal: Pratiques langagières et usages en ligne*. Ph.D. thesis, Université Grenoble Alpes (ComUE).

[16] https://lacunafund.org/

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293.

Jacques Leclerc. 2023. Sénégal. *L'aménagement linguistique dans le monde*.

Indrani Medhi, Somani Patnaik, Emma Brunskill, SN Nagasena Gautama, William Thies, and Kentaro Toyama. 2011. Designing mobile interfaces for novice and low-literacy users. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(1):1–28.

OIF. 2022. La langue française dans le monde. Technical report, Organisation International de la Francophonie.

Marie Renaudier. 2012. *Dérivation et valence en sereer. Variété de Mar Lodj (Sénégal)*. Ph.D. thesis, Université Lumière Lyon 2.

Stéphane Robert. 2022. Wolof: a grammatical sketch. In Friederike Lüpke, editor, *The Oxford guide to the Atlantic languages of West Africa*. Oxford University Press, Oxford.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2022. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2340–2354, Dublin, Ireland. Association for Computational Linguistics.

UNESCO Institute for Statistics. 2023. Education: Number of illiterates. Last consulted on 13 Feb 2024 07:49 UTC (GMT).

Daan van Esch, Tamar Lucassen, Sebastian Ruder, Isaac Caswell, and Clara Rivera. 2022. Writing system and speaker metadata for 2,800+ language varieties. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5035–5046.

World Bank. 2021. World development report 2021: Data for better lives.

Ronda Zelezny-Green, Steven Vosloo, Gráinne Conole, et al. 2018. Digital inclusion for low-skilled and low-literate people: a landscape review.

12. Language Resource References

Conneau, Alexis and Ma, Min and Khanuja, Simran and Zhang, Yu and Axelrod, Vera and Dalmia, Siddharth and Riesa, Jason and Rivera, Clara and Bapna, Ankur. 2023. *Fleurs: Few-shot learning evaluation of universal representations of speech*. 2022 IEEE Spoken Language Technology Workshop (SLT). PID https://huggingface.co/datasets/google/fleurs.

Djiba, Daouda Tandiang. 2021. *Keyword Spotting with African Languages*. Zenodo. PID https://doi.org/10.5281/zenodo.7561858.

Gauthier, Elodie and Besacier, Laurent and Voisin, Sylvie and Melese, Michael and Elingui, Uriel Pascal. 2016. *Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: a Case Study of Wolof*. European Language Resources Association (ELRA). PID https://hal.science/hal-01350037.

Nelson, Perry. 2022. *Waxal Speech Data Resources*. PID https://github.com/Waxal-Multilingual.

A. Details on speech transcription mistakes

A.1. Wolof transcriptions

Main mistakes mentioned are:

- failure to respect certain vowel lengthenings (e.g.: word "mboo**ee**m" written instead of "mboo**e**m");
- failure to respect consonant gemination internally and in the final position for a number of words (e.g.: "lo**pp**aalëb" instead of "lo**p**aalëp");
- confusion between plosives consonants in final position of certain words (/p/ *versus* /b/, /k/ *versus* /g/).

A.2. Pulaar transcriptions

Main mistakes reported in the Pulaar transcriptions are the following:

- no distinction between the consonant ɗ and the consonant ɓ (e.g.: "heeɗi" instead of "heɓi");
- use of a simple consonant instead of a prenasal consonant (e.g.: "**j**iiya" instead of "**nj**iyaa");

- concatenation of a noun and its article (e.g.: "yimɓeße" instead of "yimɓe ɓee");
- confusion in the vowel lengthening (e.g.: "d**ee**m**o**wo" instead of "d**e**m**oo**wo"; "r**ee**m**o**ße**e**" instead of "r**e**m**oo**ße ɓ**ee**").

A.3. Sereer transcriptions

In the Sereer transcriptions, phonetical, morphological and syntactical mistakes were found. Notably:

- pre-nasalised consonants ("/nd/", "/mb/", "/nj/", "/ng/") used instead of glotalised or nasal consonants ("ɓ", "ɗ", "ŋ");
- vowel lengthening not written ("ref**e**" instead of "ref**ee**","m**a**ga a mb**a**g o njirñ**a**" instead of "m**aa**ga a mb**aa**g o njirñ**aa**";
- noun and class pronoun are detached as in "xa qol axe" written "xa qola xe".

B. Speech dataset details

Rows explanation of Table 4 and Table 5:

- "Min (sec.)" is the minimum duration of an audio file in the given dataset.
- "Max (sec.)" is the maximum duration of an audio file in the given dataset.
- "Mean (sec.)" is the average duration of all the audio files in the given dataset.
- "Total audio" is the total duration of the audio set.
- "Total speech" is the total speech duration of the audio set.
- "Female speech" is the total speech duration of female speakers within the audio set.
- "Male speech" is the total speech duration of male speakers within the audio set.
- "Female speech ratio" is the percentage of speech duration of female speakers within the audio set.
- "Male speech ratio" is the percentage of speech duration of male speakers within the audio set.
- "#Turn-taking" is the number of speaker turn-takings in the whole audio set.
- "#Files" is the total number of recordings and transcriptions in the speech dataset.

"Total speech", "Female speech", "Male speech" and "#Turn-taking" durations have been computed from the Transcriber (.trs) files. This information should be treated with caution, as it depends on the accuracy of the annotations made by the transcribers. All other information in the table is calculated from audio files (.wav).

B.1. Whole set

Table 4 gives some statistics on the whole speech dataset.

Dataset statistics	**Wolof**	**Pulaar**	**Sereer**
Min (sec.)	21	20	25
Max (sec.)	3014	3033	3461
Mean (sec.)	1299	1384	1306
Total audio (hh:mm:ss)	55:11:41	31:55:10	38:12:10
Total speech* (hh:mm:ss)	51:08:50	30:06:43	36:23:37
Female speech* (hh:mm:ss)	05:12:41	04:05:07	10:12:10
Male speech* (hh:mm:ss)	45:56:09	26:01:36	26:11:26
Female speech ratio (%)	10.19	13.57	28.03
Male speech ratio (%)	89.81	86.43	71.97
#Turn-taking*	46,907	16,558	9,007
#Files	306	166	210

**extracted from annotations*

Table 4: Details about Kallaama speech dataset

B.2. Checked set

Table 5 gives some statistics on the checked subpart of the speech dataset.

Dataset statistics	**Wolof**	**Pulaar**	**Sereer**
Min (sec.)	21	117	444
Max (sec.)	2849	3033	2907
Mean (sec.)	1283	1472	1250
Total audio (hh:mm:ss)	12:49:35	11:02:28	11:06:52
Total speech* (hh:mm:ss)	11:47:34	10:56:15	10:51:33
Female speech* (hh:mm:ss)	01:27:00	01:08:09	03:12:29
Male speech* (hh:mm:ss)	10:20:33	09:48:06	07:39:03
Female speech ratio (%)	12.30	10.39	29.54
Male speech ratio (%)	87.70	89.61	70.46
#Turn-taking*	11,968	3,583	1,796
#Files	72	54	64

**extracted from annotations*

Table 5: Details about checked part of Kallaama speech dataset

C. Recordings types

The recordings are from various types of programmes and are rated on a scale of 1 to 5 based on their potential complexity for speech processing. This rating is subjective and takes into account factors such as recording duration, number of talking speakers, and recording conditions.
A rating of 1 indicates relatively low complexity, while a rating of 5 indicates relatively high complexity. This ID is the first number composing the name of the files.
Table 6 shows the number of recordings per programme type, for each language set.

Type ID	Type	Wolof	Pulaar	Sereer
1	push message	9	1	0
2	voice message	0	0	14
3	interview	22	10	15
4	radio show	120	72	67
5	focus group	2	0	9

Table 6: Number of recordings per programme type, for each language dataset

Long-Form Recordings to Study Children's Language Input and Output in Under-Resourced Contexts

Joseph R. Coffey, Alejandrina Cristia
Laboratoire de Sciences Cognitives et de Psycholinguistique,
Département d'études cognitives, ENS, EHESS, CNRS, PSL University
29 Rue d'Ulm, Paris, France 75005
jrcoffey@g.harvard.edu, alejandrina.cristia@ens.fr

Abstract

A growing body of research suggests that young children's early speech and language exposure is associated with later language development (including delays and diagnoses), school readiness, and academic performance. The last decade has seen increasing use of child-worn devices to collect long-form audio recordings by educators, economists, and developmental psychologists. The most commonly used system for analyzing this data is LENA, which was trained on North American English child-centered data and generates estimates of children's speech-like vocalization counts, adult word counts, and child-adult turn counts. Recently, cheaper and open-source non-LENA alternatives with multilingual training have been proposed. Both kinds of systems have been employed in under-resourced, sometimes multilingual contexts, including Africa, where access to printed or digital linguistic resources may be limited. In this paper, we describe each kind of system (LENA, non-LENA), provide information on audio data collected with them that is available for reuse, review evidence of the accuracy of extant automated analyses, and note potential strengths and shortcomings of their use in African communities.

Keywords: daylong recordings, voice type classification, validation, language development

1. Introduction

Technological development in the last decade has made it trivially easy to collect massive amounts of audio (and more recently, video) using wearable devices. One of the use cases in which this technology can make the biggest difference for individual and societal well-being may be in the context of early childhood education programs. Economists have argued that interventions targeting children under 3 years of age can have the greatest returns on investment (Heckman, 2008).

One crucial challenge for such interventions involves measuring the effects of such interventions, which currently entails lengthy parental interviews and/or child observations, by highly skilled individuals, making them impractical for under-resourced, multilingual contexts. In this context, long-form recordings collected with child-worn devices stand to be transformational, provided the audio(-video) data thus amassed is informative of the child's language skills and the child's environment. While speech and language technologists trusting of "state of the art" reviews thought the problem of speaker diarization was largely solved even before the advent of deep neural networks, it is now clear that even these networks crumble when faced with the formidable task of diarizing child-centered data by challenges like DIHARD (Ryant et al., 2021) and MERLION (Garcia Perera et al., 2023). And yet, through careful interdisciplinary work between speech technologists, linguists, and developmental psychologists, some progress has been made in analyzing child-centered audio to provide information about the child's speech input and output.

In this paper, we provide RAIL participants with an entry point to this emerging literature, with the dual aims of enabling both the collection of naturalistic speech data and its analysis. We first provide the background and motivation for long-form recordings. We then introduce two key hardware and software solutions that have been created and used, mainly in the fields of developmental psychology and public health. We point out both opportunities and challenges of these solutions, bearing in mind the challenges that the African context and African languages may pose.

1.1. Why and how to study young children's spoken language input and output

There is a growing interest in development economics and educational policy in how parents can positively impact their children's early development globally (UNICEF, 2019), particularly in countries where children's lives are especially vulnerable to disruption (Black et al., 2017). Many recent interventions have been aimed at increasing the frequency of parent-child conversation (Suskind et al., 2016; Weber et al., 2017; Wong et al., 2018; Ferjan Ramírez et al., 2019). Young children's early exposure to speech has been associated with language development (Hoff, 2003; Rowe, 2012; Anderson et al., 2021), school readiness (Forget-Dubois et al., 2009), and later literacy and academic

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 20–31

performance (Uccelli and Phillips Galloway, 2017).

These kinds of evaluations are difficult to conduct at scale. Researchers interested in how often children are exposed to speech must record families over long periods of time and manually transcribe the audio for speech. In their seminal "30-million-word gap" study, Hart and Risley recorded an hour of parent-child conversation every month from 42 households for 2½ years, resulting in over 1300 hours of conversation (Hart and Risley, 1995). Each hour of conversation took an estimated 8 hours to transcribe, resulting in over 10,000 man-hours of transcription. More recently, researchers working with long-form recordings estimated that accurate segmentation and transcription of children and adult speech in such data actually requires 40 hours per hour of audio data (Bergelson et al., 2023).

These methods often have limited compatibility with communities outside of urban, Western settings. They require trained numerators who have access to communities and knowledge of the local language(s) to record, transcribe, and analyze speech measures. The presence of researchers (almost always outsiders) in these communities carries a significant risk of observer effects on speech sampled. Measures of child language are also difficult to collect. Children are often raised in multilingual environments, making a single measure of language ability difficult to determine. Additionally, in communities where alloparental caregiving is common, a single parent may not be able to give a comprehensive report of children's language.

Thus, the availability of software that can quickly isolate and analyze speech from hours of recorded audio has been greatly beneficial in carrying out many of these studies. If these automated analyses were "accurate enough", long-form recordings may be particularly advantageous in characterizing the early language environments of children in Africa, especially in more rural communities. Typically, devices can be placed in children's pockets and left on for the duration of their 16-hour battery life. The devices are unobtrusive and easily forgotten, averting the discomfort created by an outside observer and providing speech estimates during the times of day and activities that a researcher normally may not have access to. Some researchers have found that these periods tend to be the most speech-dense (Casillas et al., 2019).

These systems also avoid the challenges associated with transcribing (often multiple) languages that may not have a formal writing system, or whose speakers are typically educated and literate in a different language (e.g., English, French, Arabic), a situation that is commonly encountered when working with under-resourced languages. Moreover, for many use cases, it is not necessary to produce transcripts of what was said. Instead, it is sufficient to have indicative estimates of how much children spoke and how much other people spoke, which could be (at least in theory) neutral to the specific language or languages used in the community.

2. Two systems for long-form recordings

Here we provide an overview of two examples of hardware + speech diarization systems: LENA and non-LENA alternatives (see Figure 1). The former is a widely used system developed in 2009 in the U.S. for the purpose of producing speech estimates in English, but later employed in a wide variety of settings, both urban and rural, monolingual and multilingual. The latter consists of newer systems developed in 2019 by a collaborative team of academics with the expressed goal of creating a cross-culturally robust system for producing automatic speech-based measures, encompassing a range of recording and analysis methods.

Figure 1: Top: LENA recording device placed in a child's vest (from Listen and Talk); Bottom: a USB recording device placed in the pocket of a child's shirt (from videos produced by the LAAC team)

2.1. Example 1 - The LENA System

2.1.1. Overview

LENA (Language Environment Analysis) is a combined recording and speech classification software designed for the purpose of studying children's early linguistic environments. LENA recording devices are compact (8.5cm x 5.5cm x 1.25cm) and equipped with an omnidirectional microphone, with a flat frequency response in the 20 hz-20 khz range, although the sound is bandpassed 70-10kHz (Figure 1). The LENA team often describes this as

being most sensitive to sound within a 3m radius (Ford et al., 2008), although loudness is more determinant than distance. The audio recording is eventually uploaded to the LENA software, at which point it is decompressed as 16-bit, 16kHz in PCM format.

Speech analysis techniques were developed in the early 2000s and have not been updated since the rise of recurrent neural networks. The sequence of analysis is complex and has several phases but the most relevant points are the following (see Figure 2) (Xu et al., 2008). To begin with, mel-frequency cepstral coefficients (MFCCs, representing the audio signal in a way that mimicks the human auditory system's response to different frequencies) are extracted in short windows. These are then submitted to a Minimum Duration Gaussian Mixture Model, a kind of Hidden Markov Model, to perform preliminary diarization into one of eight categories (Target Child, Male Adult, Female Adult, Other Child, Electronic Noise (i.e., TV, radio), Noise, Overlap, and Silence), each representing a distinct statistical model derived from training data. This results in a segmentation of the e.g., 16-hours of audio as a sequence of segments of each of those types, which are minimally 600 ms in length. Next, each of these segments is submitted to a likelihood ratio test to determine whether it is more likely to belong to the original category than it is to be categorized as Silence. The segments that fit better to the original category are classified as "near and clear," while the segments that do not are classified as "faint" and are excluded from subsequent analyses.

The "near and clear" adult segments are processed further to produce finer-grained estimates, using an adaptation of the CMU Sphinx phone decoder (trained on broadcast news) to estimate the number of consonants and vowels. Male and Female Adult segments are used to derive a measure of adult word counts (AWC). The "near and clear" segments attributed to the Target Child are submitted to another classifier to split the child segment into a finer sequence of speech-like, cry, and other fixed signals (e.g., snoring, burping). The speech-like sections are called "utterances" and are counted to produce a measure of child vocalizations (CVC). In addition, conversational exchanges or "turns" (CTC) between the target child and her adult caregivers are calculated by any five second interval containing a Target Child utterance and any Adult segment.

The primary objective of the LENA system is to provide users (e.g., parents, educators, researchers) with a tool for describing children's natural language environments without requiring any technical expertise nor access to computing resources. The LENA Foundation offers a variety of programs catered to the specific needs of its consumers, such as educational programming for parents (LENA Start) and educators (LENA Grow) that instruct users on how to use the recording device and software to track their own language usage around children (Elmquist et al., 2021). The LENA Foundation also offers a cloud-based processing system (LENA SP) for researchers who wish to collect and process data from multiple sites.

LENA SP is renewable subscription based service with a 5000 US$ initial setup fee. Further pricing contingent on how many concurrent participants are being tracked: 2400 US$ for up to 30 and 3900 US$ for up to 50, and 1400 US$ for each additional 25 concurrent participants. Pricing for the LENA recording devices cost 329 US$, with reductions in price for bulk purchases. LENA's recommended low-friction pocketed shirts and vests are 25$ each.

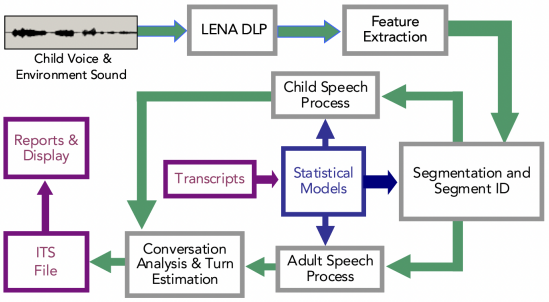


Figure 2: Illustration of the LENA audio analysis process (Xu et al., 2008)

2.2. Performance of the LENA solution

The initial validation of the LENA system was conducted by the LENA Foundation by comparing automatic speech outputs to human coded transcriptions (Gilkerson et al., 2008; Xu et al., 2009). They sampled an hour of audio each from 329 recordings of as many children between the ages of 2-42 months. Human annotators coded 10ms frames of this audio using the same categories as the LENA software. Classification was evaluated on two metrics: recall (or sensitivity) and precision. Recall measures how much of what the human annotator classified as speech LENA correctly identified, while precision measures how much of what LENA classified as speech was correctly classified. They found relatively high degrees of recall and precision for the Target Child (67% recall rate; 75% precision rate) and Female Adult categories (74% recall rate; 67% precision rate), although precision was lower (as expected) for Other Child category (64% recall rate, 27% precision rate) (Gilkerson and Richards, 2020). Subsequent studies have supported these estimates, with a review of LENA validations finding that across languages, recall and precision for cat-

egories fell 59% and 68% respectively on average (Cristia et al., 2020).

LENA has also been subject to validation in many non-English languages where it has demonstrated favorable performance. In particular, LENA outputs have been shown to perform well in tonal languages such as Shanghainese-Mandarin (Gilkerson et al., 2015) and Vietnamese (Ganek and Eriks-Brophy, 2018), as well as in languages with phonetic inventories distinct from English such as Arabic/Hebrew (Levin-Asher et al., 2023) which contain guttural consonants. These findings may bode well for studies of African languages, which are highly typologically varied and can include distinctive features such as tone (Niger-Congo languages) and click consonants (Khoisan languages) (Dryer and Haspelmath, 2013).

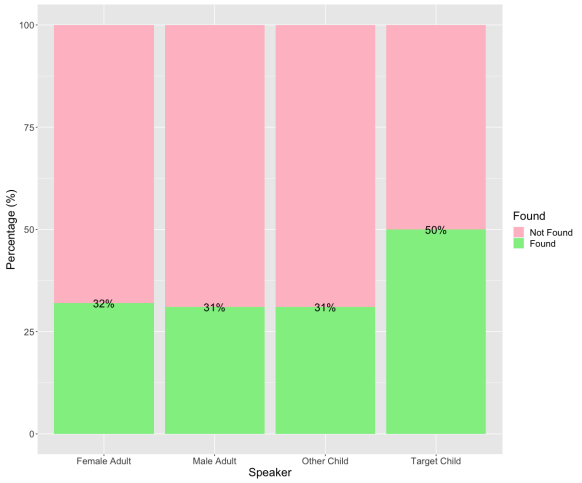
Studies have also examined correlations between transcriptions of speech and LENA speech estimates. Cristia and colleagues found high reported correlations between transcribed speech measures and adult word counts (r=0.79, n=13) and child vocalization counts (r=0.77, n=5) in their study sample, albeit lower correlations with conversational turns (r=0.36, n=6) (Cristia et al., 2020). These results suggest that LENA classification performs accurately on the majority of speech contained in recordings.

There are reasons to believe that children's language environments across African countries may be different from the samples these systems have been trained to identify, especially in rural communities. Children may spend more of their day outside, where there is more potential noise that might make speech estimates less accurate. A recent unpublished analysis, (admittedly based on very few data points), suggested no differences in accuracy across rural and urban samples (Bergelson et al., 2023).

However, Cristia and colleagues note some methodological shortcomings common to many of these studies. Firstly, most evaluations of the LENA system were not peer reviewed, and did not always fully report methods or results. Secondly, many LENA evaluations only considered audio containing speech and not Silence, Noise, or Overlap. Finally, evaluations of LENA would often focus on samples of audio containing peak instances of adult or child speech, rather than sampling randomly or periodically, which would likely have prevented noisier and more difficult to parse audio segments from being included in the evaluation. Each of these methodological choices could artificially inflate accuracy estimates.

As a follow-up to their systematic review, Cristia and colleagues examined a collection of corpora consisting of 4.6 hours of annotated English language speech from the US and UK, and 0.7 hours of speech from another corpus collected from a Tsimane' village in northern Bolivia, sampling either randomly or periodically from the audio and including non-speech categories in their evaluative metrics (Cristia et al., 2021). They found that recall rates of 50% for Target Child, but all other speaker classifications were around 30%. Precision rates were at 60% for Female Adult and Target Child, but only 43% for Male Adult and 27% for Other Child. In contrast, correlations between transcribed samples and LENA speech estimates were robust (r=0.65 for AWC; r=0.70 CVC), although CTC still lagged behind (r=0.36) (see Figure 3).

Overall, estimates of child and adult speech remained robust, but recall was markedly lower than in previous validations, and only Female Adult and Target Child retained somewhat comparable precision. As in previous validations, they found particularly poor performance distinguishing Target Child segments from Other Child segments. A recurrent finding in rural societies is that children spend much more time in conversation with other children than they do adults (Shneidman and Goldin-Meadow, 2012; Loukatou et al., 2022). As a result, systems must be able to accurately distinguish the child wearing the recording device from other children in the immediate area. To our knowledge, the LENA Foundation does not have any current plans to improve this aspect of their system.



2.3. Uses of the LENA system and available data

LENA is a flexible system, with use cases in basic research (Weisleder and Fernald, 2013; Romeo et al., 2018), early diagnosis of developmental disorders or delay (Richards et al., 2010), and early childhood intervention (Wong et al., 2018; Ferjan Ramírez et al., 2019; Elmquist et al., 2021).

In general, most studies using LENA have come from the U.S. where the LENA Foundation is based

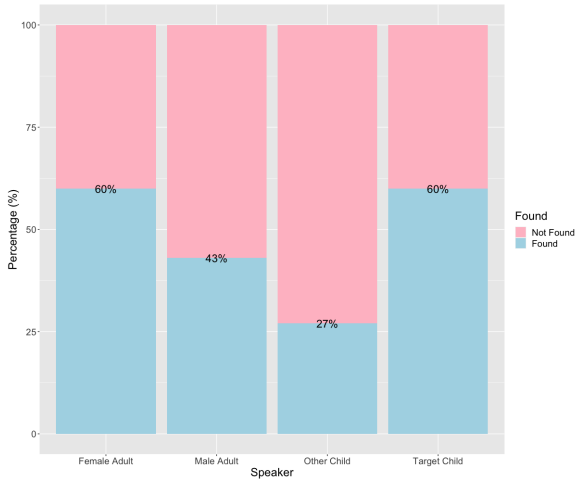


Figure 3: Recall (above) and precision (below) statistics from recordings of US, UK, and Tsimane households (Cristia et al., 2021)

(Wang et al., 2020). Homebank, a publicly accessible repository of long-form recordings, contains 18 corpora of recordings from over 300 children. Of these, four contain data from languages other than English, three of which were collected outside of the U.S. However, LENA is seeing increasing use internationally. A recent multi-site study examined LENA use across 12 countries in 10 different languages, including three rural communities (Tsimane' in Bolivia, Yélî Dnye on Rossel Island, and Wolof-speaking children in rural Senegal) (Bergelson et al., 2023).[1]

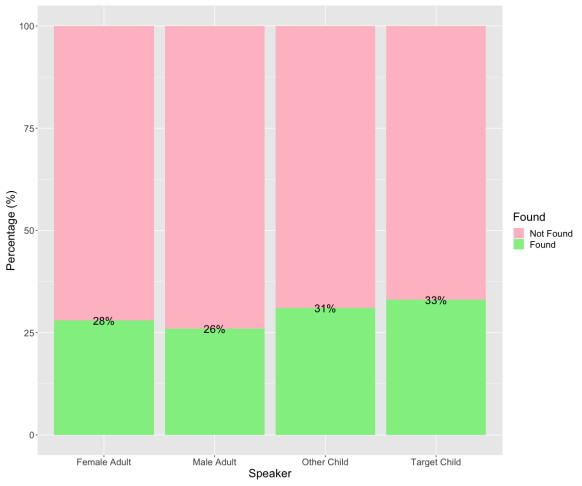
2.4. Feasibility of use in African countries

The accuracy of speech diarization systems is contingent on their ability to address the particular challenges of rural African communities. As of yet, there has only been a single evaluation of the LENA system conducted in an African language to our knowledge. Coffey, Zhang, & Spelke examined 52 hours of audio from a small sample of 4 Akan-speaking children (15.5 to 41mos) living in Accra, Ghana (Coffey et al., 2023). They sampled 2 minutes of audio periodically from every hour of recording and coded each according to the ACLEW coding scheme (Cristia et al., 2021). They found relatively low rates of Recall across all speakers (28% of Female Adult; 26% of Male Adult; 31% of Other Child and 33% of Target Child). They also found higher rates of Precision for Female Adult (45%) and Target Child (56%), but not for Male Adult (32%) or Other Children (13%) (Figure 4).

[1]This data is available for reuse through the EL1000 corpus via GIN: https://gin.g-node.org/LAAC-LSCP/EL1000

Comparing these findings to those illustrated in Figure 3, we find similar rates of recall across all speaker categories except for Target Child, which are lower (33% vs. 50%). In contrast, they find comparable rates of precision for Target Child, but somewhat lower rates for all other categories. These results suggest that LENA accuracy may be lower in noisier settings (only 10% of Cristia et al.'s sample was drawn from a rural non-Western sample), but it may capture comparable amounts of speech to other similar studies. LENA also appears to experience difficulty distinguishing Target Child from Other Child speech: 25% of human coded Target Child speech was classified as Other Child by the LENA device.

Likewise, there has only been a single published study in Africa that has related LENA speech measures to children's language. Weber, Fernald, and Diop assessed the impact of a parenting intervention designed to encourage more verbal engagement between mothers and their 4- to 31-month old children in rural Senegal by tracking child-directed speech throughout the day using LENA (Weber et al., 2017). They found children of mothers who received the intervention had larger vocabularies than children of controls. Despite the effectiveness of the intervention in increasing maternal speech during a short recorded play session, they did not find LENA speech measures to be correlated with outcomes in either group. This finding is at odds with results from studies of LENA elsewhere, which have found consistent correlations between LENA speech measures and children's language roughly equivalent in size to studies using transcribed speech measures (Wang et al., 2020; Anderson et al., 2021).



2.5. Summary

The principal advantages of the LENA system are its popularity, ease of use, availability of data, and rigorous validation across multiple languages by an

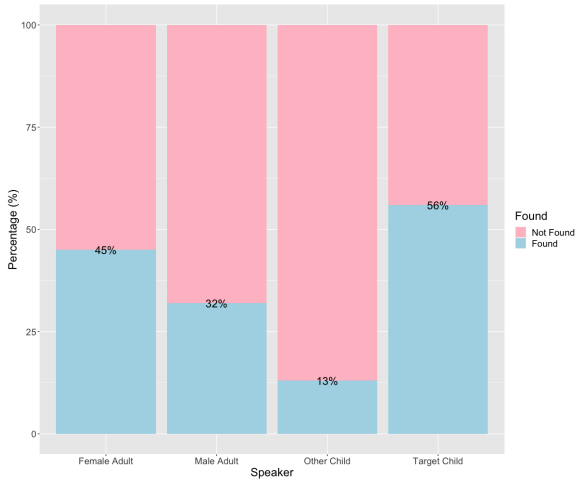


Figure 4: Recall (above) and precision (below) statistics from recordings of Ghanaian households (Coffey et al., 2023)

increasingly international body of users. LENA is an effective speech diarization system that has promising applications in research, education, and public health in Africa. However, there are still significant shortcomings. The LENA SP is expensive (minimum 7400 US$, not including the cost of hardware), making implementation difficult with low-budgeted local projects, as well as at scale. The system, including hardware and software, is also proprietary, making individual alterations or improvements for specific projects impossible to be implemented. Because LENA SP holds data on cloud servers hosted within the U.S., users in other countries may find it difficult to use LENA without violating data privacy laws. Finally, LENA has been shown to have low accuracy distinguishing Other Children from the Target Child, which may create problems in communities where child caregiving is common (Barry and Paxson, 1971; Zukow-Goldring, 2002) and most speech to children comes from their siblings and peers (Shneidman and Goldin-Meadow, 2012; Loukatou et al., 2022). While there are many advantages to using LENA in projects with sufficient budgeting and institutional approval, these factors may make using LENA impractical in other contexts.

3. Example 2 - Non-LENA

3.1. Overview

Researchers who were unable or unwilling to use the LENA system have turned to other recorders. For example, Marisa Casillas fit a baby-sized harness with an Olympus recorder (initially produced for linguistic work on conversations), and used it to collect long-form data in a Tseltal village in Mexico and several other locations in Rossel Island, Papua New Guinea (Casillas et al., 2019, 2021). Cristia and colleagues used this Olympus as well as even smaller, "spy" USB devices in Bolivia and Vanuatu (Scaff et al., 2024; Cristia et al., 2023). The USB-based method attracted considerable attention from economists working in the Pacific area because its low price (20 US$/device) enabled data collection at scale. The precise technical characteristics of the microphones, recorders, and sound files depend on the specific equipment and its settings. For example, the Olympus Casillas used can be set to record .mp3 files, in which case a battery would last for 22 consecutive hours, with frequencies up to 22 kHz but lower bit rates than if .wav is used instead. The "spy" USBs often record with a sampling frequency closer to LENA's (15kHz) and 8-bit depth.

Once recordings are obtained with any relevant device, they can be processed using an open-source software called the Voice Type Classifier (VTC). The key aspects of VTC were developed during and after the Jelinek Summer Workshop on Speech and Language Technology, which allowed testing a variety of input features, tasks, and architectures (Lavechin et al., 2020). The best model received as input the raw waveforms and processed them through a Sincnet, followed by a Long Short-Term Memory (LSTM) neural network with three fully connected layers (see Figure 5). The Sincnet is a type of neural architecture that attempts to learn audio features to describe the input signal it is given, and in our experiments, we found it outperformed other forms of representation (like the MFCCs used by LENA). LSTMs are a type of recurrent neural networks, particularly suited to sequential data, which is appropriate for a time series like speech. Through this process, the audio is diarized into Female Adult, Male Adult, Target Child, and Other Children, with any of these overlapping with the others. The training set contained child-centered data from various linguistic backgrounds and environments including languages like Min (a Sino-Tibetan tonal language), French, Ju'hoan (a Khoisan language with clicks), Tsimane' (an indigenous Bolivian language), and English, covering both urban and rural settings, as well as multilingual contexts. This broad training was aimed to enhance VTC's ability to generalize to new datasets, which is particularly useful for researchers working in under-resourced language contexts.

ALICE, an open-source reusable software, was developed to return word, syllable, and phoneme counts in VTC-identified male and female adult vocalizations (Räsänen et al., 2021). The pre-trained version of ALICE that was released to be applied to any language employs SylNet, an end-to-end neural network syllable count estimator, together with

Figure 5: Voice type classifier architecture, illustrated on 2s of input audio waveform (Lavechin et al., 2020)

signal-level features (such as utterance duration, total energy, and zero-crossings), plus fixed weights from a linear regression (jointly fit for 7 corpora, including American and British English, Tseltal, Yélî Dnye, and Argentinean Spanish) to provide estimates of word, phoneme, and syllable counts. At present, it only does this for adult speech, and not for the key child or other children's speech. The challenge for applying it key child speech is finding sufficient quantities of transcribed data. For vocalizations attributed to other children, an additional challenge is the heterogeneity of such a category, covering speech by infants all the way to pre-pubescent children.

A deep-learning, open-source solution has also been proposed to detect infant crying (Yao et al., 2022). In a nutshell, a support-vector machine (SVM) classifier was trained using a combination of acoustic features and deep spectrum features extracted from a customized version of the AlexNet architecture (comprising five convolutional layers and three fully connected layers), with adjustments to the input and output layers to accommodate the data.

Using a non-LENA device and software requires a smaller budget than LENA, provided that technical knowledge and computational resources are not taken into account. For example, to compare with LENA's cheapest option, one could purchase 30 USB devices and give one to each of 30 families, to record their child with monthly. Including only the devices, this would require a budget of about 600 US$, which is the price of 2 LENA devices. This hardware is also more likely to be available within the country, whereas LENA devices must be ordered from the U.S.

In contrast, if technical knowledge and computational processing are taken into account, we doubt that costs would be much lower for this option than LENA's, although one would have to run experiments to be certain. Unlike LENA, which can be used by anyone who can handle a GUI and a web browser, all the non-LENA options require more technical knowledge and access to resources. For instance, for VTC (and ALICE, which depends on VTC), it is necessary to install *pyannote* (Bredin et al., 2020) and all of its dependencies, and to know how to create a *conda* environment. As for resources, although we know of researchers who were able to install it and analyze audio-recordings on a mac laptop, VTC would ideally be ran in a GPU, where one can benefit from analyses requiring 1/45 of the recording time (versus 1/4 in CPU). One option researchers have used is to create an AWS instance to run the analyses (Peurey et al., 2024), in which case the cost of running both VTC and ALICE was estimated as 0.20 US$ per hour of audio analyzed (so about 3 US$ per 15-hour recording). We do not know of similar estimations for the cry detection system.

3.2. Performance of the non-LENA software

Each of the three open-source solutions has been benchmarked against LENA, and shown to outperform or match the performance of the corresponding step in LENA software. Since LENA software can only be applied to audio collected using the LENA device, these evaluations reflect performance for the software holding device constant.

For voice type classification, VTC outperformed LENA software for all categories in an evaluation that was based entirely on English urban child-centered data. In terms of F-score, performance was: 69% versus 55% for the target child; 33% versus 29% for other child; 63% versus 43% for female adult; and 43% versus 37% for male adult. See (Lavechin et al., 2020) for details. We point out that, although outperforming LENA, VTC's performance for Other Child is far from reasonable: 33% means that most of the time, the system gets this category wrong. In unpublished work, the team that developed VTC has looked for improvements without compromising performance in the other categories by increasing the amount of data from this category. Indeed, they noticed that the original training dataset had a good representation of female adult voices (46 hours) and target child (34 hours), whereas male adults (1 hour) and other children (4 hours) were rarer in those data. The team thus targeted human annotations in families where there were siblings, increasing the representation of other child to 4.5 hours. However, this did not suffice to improve performance. Annotation efforts are still ongoing, but this is slow work as this type of challenging data requires about 40 minutes of work to segment one minute of data, and often it is necessary to employ even more time and effort to come to learn the individual children's voices. A key challenge with the other child category is that, unlike the key child, it is not a homogeneous category, applying to a single individual. Thus, it covers any child, from pre-linguistic babies all the way to 13-year-olds. Breaking it into subcategories by age did not seem promising given the amount of data available. Thus, this remains a challenging

problem.

For word counting, two types of analyses were reported by Räsänen and colleagues, which also was based on English urban child-centered data. One compared correlations in the total counts over 2-minutes of audio across the two softwares, which is similar to the majority of work evaluating LENA accuracy. ALICE outperformed the LENA software in two out of 4 corpora (correlations between human and automated word counts around .9 for ALICE, .75-.8 for LENA); was similar for a third one (correlations around .8 for both); and under-performed for the last one (correlations .65 for ALICE, .7 for LENA). However, the authors argue that sometimes it is not sufficient to rely on correlations, since the algorithms may over- or under-estimate word counts. They therefore report a second metric, the median of the absolute error rate, which is less forgiving. This metric showed an advantage for ALICE across the board, with error rates 20% higher for LENA than ALICE in all 4 corpora (Räsänen et al., 2021).

For cry detection, Micheletti and colleagues similarly report correlations and error rates in terms of the number of cries discovered, using as test set English urban child-centered data. Considering 5-minutes, which is a common unit in previous work evaluating LENA, the two algorithms were quite matched in their performance, with correlations around .79 for Yao's DL algorithm and .75 for the LENA software. However, LENA severely underestimated total duration, underestimating cry duration by about 51 minutes per 24h of audio, versus the open source alternative's slight over-estimation of 35 seconds per 24h of audio (Micheletti et al., 2023).

These results are not surprising given that the LENA software relies on outdated input features and technology. Two important caveats are in order. First, since the above evaluations were done by the same teams who proposed the open source tools, there could be a conflict of interest. Moreover, typically those evaluations covered a small number of languages and settings, whereas there have been many more independent evaluations of the LENA solution. Second, and most importantly, evaluations always benchmarked against LENA, which entailed using audio collected with LENA hardware and on English urban child-centered data. These results may not generalize to other recording devices and/or languages and settings, an issue that should be addressed in future work. Interestingly, an informal evaluation suggests that devices other than LENA's can result in higher accuracy for talker diarization when using VTC (LAAC-LSCP).

3.3. Use of the non-LENA system and available data

The vast majority of previous work has opted for the LENA solution, and thus only a handful of studies have been published with the alternative. Setting aside technical contributions, there are to our knowledge only five published or public studies, four relying on manual annotation (Casillas et al., 2019, 2021; Scaff et al., 2024; Bunce et al., 2020), and one on automated analyses (Cristia et al., 2023). None of these data have been made available for reuse yet.

3.4. Feasibility of use in African countries

We do not know of any work that has employed a non-LENA alternative in Africa. However, Alex Cristia has obtained funding to help support researchers interested in employing the non-LENA system by lending them equipment and expertise, provided that goals are compatible with the project "Experience effects in early language."[2]

3.5. Summary

Overall, the combination of affordable hardware and advanced software tools provides researchers with a valuable means to explore vocalization data across diverse linguistic contexts, offering insights into child development and linguistic diversity. Because these solutions were created with cross-cultural work in mind (training data from non-English and non-U.S. settings, affordable open-source tools, flexible hardware choice) they may be better suited for work in African communities. However, due to the majority of long-form recording studies being conducted with LENA, there is not as much published evidence that these devices provide as accurate estimates in noisier non-English settings (although its training data includes this sort of audio), nor is there as much publicly available data. LENA's ease of use and institutional support from the LENA Foundation may also make its use more feasible for parties less familiar with these kinds of tools.

4. Conclusion

In this paper, we described two kinds of systems for collecting and analyzing long-form recordings of children's early language environments. We reviewed each of their underlying audio processing systems, compared their validity across settings

[2]More information can be obtained on https://exelang.fr/call-for-data.

and languages, and outlined the potential advantages and disadvantages of their use in African settings.

The first system, LENA is a combined daylong recording and analysis system developed by the LENA Foundation, based on data drawn primarily from the U.S., that uses a Gaussian mixture model approach for segmenting audio by source/speaker and producing estimates of speech from children and adults. The second kind of system, non-LENA approaches, uses an open source program (VTC) based on a neural network architecture to diarize speech from audio collected from many different possible devices (e.g., Olympus recording devices, "spy" USB recorders). These segments can then be input into further speech processing algorithms (e.g., ALICE) to derive estimates of speech.

Our review suggests that the principal advantages of using LENA are its ease of use, support, and widespread adoption. The LENA devices and software are designed to be intuitive and easily understood. The LENA Foundation also provides institutional support, from project advisement to cloud computing services. For this reason, LENA has become a popular tool in research and education, and has undergone validation in many different languages and countries. Data from many of these studies are also publicly available. However, LENA and its hardware can be prohibitively expensive. Data hosting may also be difficult depending on the country research is done in. Finally, LENA is a proprietary system, and thus neither the software nor the hardware can be changed, updated, or adapted for use in specific contexts.

In contrast, non-LENA solutions are cheap, flexible, and based on up-to-date technical methods designed with cross-linguistic work in mind. There is a growing network of researchers using these tools and contributing directly to their continued development. Hardware can be adjusted as needed, and algorithmic methods for speech analysis are constantly being updated. But due to the newness of these systems, there is not currently a large user base, nor the same degree of validation as the LENA system has. There is also no publicly available data using these methods. These solutions also require more technical knowledge to use and support is more limited than what LENA provides.

Overall, we found that very little work has been done in Africa with either of these systems. In addition, we found similar shortcomings for both solutions. Namely, both systems have been found to perform poorly distinguishing speech from the target child from speech from other children, and while the community developing non-LENA solutions aims to address this challenge, this work is still very much ongoing. This is a potential obstacle to the analysis of speech drawn from naturalistic contexts in many African communities, where children are most exposed to speech from their siblings and peers on a daily basis (Loukatou et al., 2022). Despite these challenges, long-form recordings have applications in Africa that have the potential to be highly impactful for research, early childhood education, and public health. Thus, it is our hope that researchers, educators, and policymakers consider their use.

5. Acknowledgements

We thank: the J. S. McDonnell Foundation Understanding Human Cognition Scholar Award; European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095).

6. Ethics Statement

Long-form recording systems have many promising applications in research, education, and public health in Africa. However, there are also ethical considerations inherent to the collection of sensitive data from African communities. In particular, it is important to be aware of the risk of bias, on the part of both the researcher and the algorithm itself. These biases can affect how data is interpreted and acted upon, which could have unintended consequences for these communities. It is also important that consent is obtained in a way that is both culturally appropriate and in line with local and national privacy laws. Finally, the benefits of research should be determined in collaboration with local communities and distributed fairly. For further discussion of ethical considerations, see (Léon et al., 2024).

7. References

N.J. Anderson, S.A. Graham, H. Prime, J.M. Jenkins, and S. Madigan. 2021. Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development*, 92(2):484–501.

H. Barry and L.M. Paxson. 1971. Infancy and early childhood: cross-cultural codes. *Ethnology*, 10(4):466–508.

E. Bergelson, M. Soderstrom, I. C. Schwarz, C. F. Rowland, N. Ramírez-Esparza, L. R. Hamrick, E. Marklund, M. Kalashnikova, A. Guez, M. Casillas, L. Benetti, P. van Alphen, and A. Cristia.

2023. Everyday language input and production in 1,001 children from six continents. *Proceedings of the National Academy of Sciences*, 120(52):e2300671120.

M.M. Black, S.P. Walker, L.C. Fernald, C.T. Andersen, A.M. DiGirolamo, C. Lu, D.C. McCoy, G. Fink, Y.R. Shawar, J. Shiffman, A.E. Devercelli, Q.T. Wodon, E. Vargas-Baron, and S. Grantham-McGregor. 2017. Early childhood development coming of age: Science through thelife course. *Lancet*, 389(10064):77–90.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. IEEE.

J. Bunce, M. Soderstrom, E. Bergelson, C. Rosemberg, A. Stein, M. Migdalek, and M. et al. Casillas. 2020. A cross-cultural examination of young children's everyday language experiences. *PsyArxiv*.

M. Casillas, P. Brown, and S. C. Levinson. 2021. Early language experience in a papuan community. *Journal of Child Language*, 48(4):792–814.

M. Casillas, P. Brown, and S.C. Levinson. 2019. Early language experience in a tseltal mayan village. *Child Development*, 91(5):1819–1835.

J. Coffey, S. Zhang, and E. Spelke. 2023. Validation of lena measures of parent speech in ghana. In *Proceedings of the Society for Research in Child Development 2023 Biennel Meeting*, Salt Lake City, Utah.

A. Cristia, F. Bulgarelli, and E. Bergelson. 2020. Accuracy of the language environment analysis system segmentation and metrics: A systematic review. *Journal of Speech, Language, and Hearing Research*, 63:1093–1105.

A. Cristia, L. Gautheron, and H. Colleran. 2023. Vocal input and output among infants in a multilingual context: Evidence from long-form recordings in vanuatu. *Developmental Science*, 26(4):e13375.

A. Cristia, M. Lavechin, C. Scaff, M. Soderstrom, C. Rowland, O. Räsänen, and J. Bunce. 2021. A thorough evaluation of the language environment analysis (lena) system. *Behavior Research Methods*, 53:467–486.

M.S. Dryer and M. (eds.) Haspelmath. 2013. *World Atlas of Language Structures Online (v2020.3)*. Zenodo.

M. Elmquist, L.H. Finestack, A. Kriese, E.M. Lease, and S.R. McConnell. 2021. Parent education to improve early language development: A preliminary evaluation of lena start. *Journal of Child Language*, 48(4):670–698.

N. Ferjan Ramírez, S.R. Lytle, M. Fish, and P.K. Kuhl. 2019. Parent coaching at 6 and 10 months improves language outcomes at 14 months: A randomized controlled trial. *Developmental Science*, 22:e12762.

M. Ford, C.T. Baer, D. Xu, U. Yapanel, and S. Gray. 2008. Audio specifications of the dlp-0121. LTR 03-2, LENA Foundation, Boulder, CO.

N. Forget-Dubois, G. Dionne, J.P. Lemelin, D. Pérusse, R.E. Tremblay, and M. Boivin. 2009. Early child language mediates the relation between home environment and school readiness. *Child Development*, 80(3):736–749.

H.V. Ganek and A. Eriks-Brophy. 2018. A concise protocol for the validation of language environment analysis (lena) conversational turn counts in vietnamese. *Communication Disorders Quarterly*, 39(2):371–380.

L.P. Garcia Perera, Y.H.V. Chua, H.X. Liu, F.T. Woon, A.W.H. Khong, J. Dauwels, S. Khudanpur, and S.J. Styles. 2023. Merlion ccs challenge evaluation plan. *PsyArxiv*.

J. Gilkerson, K.K. Coulter, and J.A. Richards. 2008. Transcriptional analyses of the lena natural language corpus. LTR 06-2, LENA Foundation, Boulder, CO.

J. Gilkerson and J.A. Richards. 2020. A guide to understanding the design and purpose of the lena® system. LTR 12, LENA Foundation, Boulder, CO.

J. Gilkerson, Y. Zhang, D. Xu, J.A. Richards, X. Xu, F. Jiang, J. Harnsberger, and K. Topping. 2015. Evaluating language environment analysis system performance for chinese: A pilot study in shanghai. *Journal of Speech, Language, and Hearing Research*, 58(2):445–452.

B. Hart and T.R. Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

J. Heckman. 2008. Schools, skills, and synapses. *Economic Inquiry*, 46(3):289–324.

E. Hoff. 2003. The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, 74(5):1368–1378.

LAAC-LSCP. Longform hardware audio test repository.

M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia. 2020. An open-source voice type classifier for child-centered daylong recordings. In *Proceedings of Interspeech 2020*. ISCA.

B. Levin-Asher, O. Segal, and L. Kishon-Rabin. 2023. The validity of lena technology for assessing the linguistic environment and interactions of infants learning hebrew and arabic. *Behavior Research Methods*, 55(3):1480–1495.

G. Loukatou, C. Scaff, K. Demuth, A. Cristia, and N. Havron. 2022. Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, 49(6):1173–1192.

Meera Léon, M., S.S., A.C. Fiévet, and A. Cristia. 2024. Long-form recordings in low-and middle-income countries: recommendations to achieve respectful research. *Research Ethics*, 20(1):96–111.

M. Micheletti, X. Yao, M. Johnson, and K. de Barbaro. 2023. Validating a model to detect infant crying from naturalistic audio. *Behavior Research Methods*, 55:3187–3197.

L. Peurey, W. N. Havard, X. N. Cao, and A. Cristia. 2024. Full description of an automated pipeline for providing personalized feedback based on audio samples. *Center for Open Science*, b2746.

O. Räsänen, S. Seshadri, M. Lavechin, A. Cristia, and M. Casillas. 2021. Alice: An open-source tool for automatic measurement of phoneme, syllable, and word counts from child-centered daylong recordings. *Behavior Research Methods*, 53:818–835.

A. Richards, D. Xu, and J. Gilkerson. 2010. Development and performance of the lena automatic autism screen. LTR 10-1, LENA Foundation, Boulder, CO.

R.R. Romeo, J.A. Leonard, S.T. Robinson, M.R. West, A.P. Mackey, M.L. Rowe, and J.D. Gabrieli. 2018. Beyond the 30-million-word gap: Children's conversational exposure is associated with language-related brain function. *Psychological Science*, 29(5):700–710.

M. Rowe. 2012. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, 83(5):1762–1774.

N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman. 2021. The third dihard diarization challenge. In *Proceedings of Interspeech 2021*, pages 3570–3574. ISCA.

C. Scaff, M. Casillas, J. Stieglitz, and A. Cristia. 2024. Characterization of children's verbal input in a forager-farmer population using long-form audio recordings and diverse input definitions. *Infancy*, 29:196–215.

L. A. Shneidman and S. Goldin-Meadow. 2012. Language input and acquisition in a mayan village: How important is directed speech? *Developmental Science*, 15(5):659–673.

D.L. Suskind, K.R. Leffel, E. Graf, M.W. Hernandez, E.A. Gunderson, S.G. Sapolich, E. Suskind, L. Leininger, S. Goldin-Meadow, and S.C. Levine. 2016. A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of Child Language*, 43(2):366–406.

P. Uccelli and E. Phillips Galloway. 2017. Academic language across content areas: Lessons from an innovative assessment and from students' reflections about language. *Journal of Adolescent and Adult Literacy*, 60(4):395–404.

UNICEF. 2019. *For Every Child, Every Right: The Convention on the Rights of the Child at a crossroads*. United Nations Children's Fund (UNICEF), New York.

Y. Wang, R. Williams, L. Dilley, and D. M. Houston. 2020. A meta-analysis of the predictability of lena™ automated measures for child language development. *Developmental Review*, 57:100921.

A. Weber, A. Fernald, and Y. Diop. 2017. When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural senegal. *Child Development*, 88(5):1513–1526.

A. Weisleder and A. Fernald. 2013. Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, 24(11):2143–2152.

K. Wong, M. Boben, and M. C. Thomas. 2018. Disrupting the early learning status quo: Providence talks as an innovative policy in diverse urban communities.

D. Xu, U. Yapanel, and S. Gray. 2009. Reliability of the lena language environment analysis system in young children's natural home environment. LTR 05-2, LENA Foundation, Boulder, CO.

D. Xu, U. Yapanel, S. Gray, and C.T. Baer. 2008. The interpreted time segments (its) file. LTR 04-2, LENA Foundation, Boulder, CO.

X. Yao, M. Micheletti, M. Johnson, E. Thomaz, and K. de Barbaro. 2022. Infant crying detection

in real-world environments. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135. IEEE.

P. Zukow-Goldring. 2002. Sibling caregiving. In *Handbook of parenting: Being and becoming a parent (2nd Edition)*, pages 253–286, Mahwah, NJ. Lawrence Erlbaum Associates Publishers.

Developing Bilingual English-Setswana Datasets for Space Domain

Tebatso G. Moape[1], **Sunday O. Ojo**[2], **Oludayo O. Olugbara**[3]
University of South Africa[1], Durban University of Technology[2,3]
28 Pioneer Ave, Florida Park, 1704, South Africa[1],
43 M L Sultan Rd, Greyville, Durban, 4001, South Africa[2,3]
moapetg@unisa.ac.za[1], sundayo1@dut.ac.za[2], oludayoo@dut.ac.za[3]

Abstract

In the current digital age, languages lacking digital presence face an imminent risk of extinction. In addition, the absence of digital resources poses a significant obstacle to the development of Natural Language Processing (NLP) applications for such languages. Therefore, the development of digital language resources contributes to the preservation of these languages and enables application development. This paper contributes to the ongoing efforts of developing language resources for South African languages with a specific focus on Setswana and presents a new English-Setswana bilingual dataset that focuses on the space domain. The dataset was constructed using the expansion method. A subset of space domain English synsets from Princeton WordNet was professionally translated to Setswana. The initial submission of translations demonstrated an accuracy rate of 99% before validation. After validation, continuous revisions and discussions between translators and validators resulted in a unanimous agreement, ultimately achieving a 100% accuracy rate. The final version of the resource was converted into an XML format due to its machine-readable framework, providing a structured hierarchy for the organization of linguistic data.

Keywords: Digital language resources, Setswana bilingual dataset, Space domain translation

1. Introduction

The Princeton Wordnet (PWN) is an English lexical database formally introduced by Miller (1995) and developed at Princeton University. It has served as the primary lexical semantic resource for numerous researchers in the field of Natural Language Processing (NLP) and computational linguistics (Batsuren et al., 2019). The presence of this resource has facilitated the development of various NLP applications such as machine translation, information retrieval, and tools for word sense disambiguation. Additionally, the availability of PWN has provided researchers with the capability to evaluate and compare the effectiveness of various language models and applications.

However, languages such as Setswana face a scarcity of resources (Sebolela, 2009, Marivate et al., 2020), resulting in limited availability of linguistic tools and applications. Furthermore, the current tools are frequently created within isolated projects, each with curated data tailored to its particular scope. This fragmentation poses a challenge for researchers in effectively collaborating and comparing their work.

Apart from unannotated parallel-aligned corpora and word list dictionaries extracted from government websites, the only available resource comparable to the PWN is the African Wordnet (AWN). The AWN project was initiated with the aim of promoting multilingualism and facilitating the development of language tools and resources for South African (SA) languages (Bosch and Griesel, 2017). Currently, wordnets have been developed for Setswana, isiXhosa, isiZulu, Sesotho sa Leboa, and Tshivenda. The AWN holds significance due to the scarcity of data in South African languages. This makes the AFW a crucial resource.

In an effort to contribute to the development of resources for SA languages in general and Setswana in particular, this paper presents a Setswana lexicon. The lexicon was developed by translating a subset of the PWN through expert translation, expansion, and domain adaptation methods. The chosen domain for the translation focused on space-related concepts. The outcome of this project is a Setswana lexicon comprising of 6016 synsets, with lemmas, glosses, and usage examples. The use of expert-driven translation was to ensure the generation of high-quality translations, and the decision to focus on a specific domain was made to enable the Setswana lexicon's relevance and applicability in the targeted context.

The rest of the paper is structured as follows: section 2 presents relevant literature related to the development of semantic resources across languages and the state of the art of the available language resources for Setswana. Section 3 outlines the techniques and methodologies used for the resource development. Resource evaluation and results are presented in section 4. Section 5 concludes the paper.

2. Related Works

This section is divided into two subsections. The first sub-section focuses on relevant literature related to the development of semantic resources across languages. This literature provides a foundation and context for the methodology employed in developing the resource presented in this paper. The second sub-section provides a

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 32–36

high-level overview of available resources for Setswana.

2.1 Development of Language Resources

Monolingual lexicons are constructed using two methods, namely, the expansion method and the merge method (Bosch and Griesel, 2018). In the expansion method, developers translate a subset of English synsets from PWN. The merge method involves the creation of synsets for the target languages, which are then merged with PWN synsets. The key distinction between the two methods is that the expansion method results in the target language inheriting PWN's semantic structure, while the merge method entails the creation of a new semantic structure for the target language, which is subsequently merged with PWN's semantic network. Resources conducted using these methods include (Batsuren et al., 2019, Bella et al., 2020).

In focusing on the space domain, this study used the expansion method for Setswana. The rationale behind this choice stems from the existence of similar field concepts within both English and Setswana space domains. To ensure that all the Setswana space concepts were included, concepts present in Setswana but absent from the translated set were added to the dataset and subsequently lexicalized into English. This guarantees a comprehensive coverage of space-related concepts in Setswana.

2.2 State of the Art on Setswana Resources

The importance of the availability of language resources cannot be overstated, as they play a crucial role in the preservation of languages and serve as an enabler for the advancement and development of NLP tools. In efforts to create, consolidate, and disseminate language resources for diverse SA languages, the South African Centre for Digital Language Resources (SADiLaR), in collaboration with various universities was established for this purpose. Supported by the Department of Science and Innovation (DSI), SADiLaR plays a significant role in facilitating the centralization of these language resources, contributing to their accessibility and use. This section outlines the text resources accessible for Setswana on SADiLaR, providing an overview of the presently available resources accessible to researchers.

In summary, currently, including the AFW (Bosch and Griesel, 2017), there is a total of 24 various types of text corpora. This includes multilingual word and phrase translations, phrase chunk annotated corpus, monolingual corpora, test suite, data treebank, named entity annotated corpus, annotated text corpora, and English-Setswana parallel corpora (Lastrucci et al., 2023, McKellar and Puttkammer, 2020). There is also domain-based data where English data from specific domains were translated into multiple SA languages, including Setswana. This encompasses data from domains such as soccer, mathematics, technology, health, natural sciences, arts and culture, government, elections, and parliamentary proceedings. The dataset presented in this paper specifically falls within the space domain, further expanding the scope of available data resources for the Setswana language.

3. Methodology

This paper adopted the expert-sourced expand approach to develop the presented resource. A subset of words, glosses, and examples from the PWN were translated and validated by Setswana language experts. The methodology consists of four phases, namely, translation data generation, translation, reformatting, and validation. The translations were conducted on a Microsoft Excel Spreadsheet. The following sub-sections expand on the structure of the spreadsheet, data generation, translation, validation, and reformatting phases.

3.1 Translation via Microsoft Excel Spreadsheets

The Microsoft Excel Spreadsheet contains source and target lemmas, synsets, glosses, and examples. These are defined as:

3.1.1 Lemma

A lemma is the canonical form (dictionary form, citation form) of a set of words (word forms). For example, *tsamaya (go)* is the lemma of the words *tsamaya (go)*, *wa tsamaya(goes)*, and *o tsamaile (went)*.

3.1.2 Synsets

A synset is a set of synonyms that represent a single concept or idea in linguistics which consists of lemmas. Each synset represents a unique concept, and words within the same synset are considered synonymous with one another. Synsets provide a way to organize and understand the relationships between words and their meanings in a structured format.

3.1.3 Gloss

A text or sentence that describes the concept, i.e., a lemma.

3.1.4 Example

A text or sentence(s) that clarify the exact meaning of the described concept. Examples are also used to clarify and demonstrate how the lemma/concept is used in a sentence.

3.2 Translation Data Generation

For the translation data generation phase, a domain-specific adaptation method was used.

This method focuses on the creation of language resources based on a specific domain or subject area. The following criteria were used when selecting the subset of English synsets to be translated.

3.2.1 Domain Identification

The space domain dataset was selected.

3.2.2 Data Extraction

Lemmas, glosses, and examples were extracted and transferred to an Excel file.

Wordnet data is normally divided into four parts of speech categories, nouns, verbs, adjectives, and adverbs. To narrow the focus, this study focused on data in these four categories that are in the space domain.

3.3 Translation

The translations were conducted on a Microsoft Excel file. The Excel sheet contains a number of synsets in the source language to be translated into the target language. In this case, English was the source language, and Setswana was the target language. The file is organized in a pair-wise format (source language column - target language column). The translator fields consist of the following:

3.3.1 Synset lemmas columns

Column C: Contains a comma-separated list of lemmas of the source language.

Column D: The translator provides a comma-separated list of the synset lemmas of the target language.

3.3.2 Synset gloss columns

Column F: Contains the synset gloss in the source language.

Column G: The translator provides the synset gloss in the target language.

3.3.3 Synset examples columns

Column I: Contains the synset examples in the source language.

Column J: The translator provides the synset examples in the target language.

3.3.4 Translator notes columns

Column L: The translator can provide notes related to the synset translation if there are any.

3.4 Validation

The same Excel sheet used for translation was used for validation. The validator fields consist of the following:

3.4.1 Synset lemmas validation

The validator provides his validation on the lemmas in the target language in column E. The validator can choose between:

- Accepted: If the validator finds that the lemmas are complete and do not contain any errors such as spelling errors, she/he writes A (for accepted).

- Rejected: If the validator finds that the lemmas are not correct, or there are missing lemmas, or lemmas that do not belong to the synset, she/he writes R (for rejected) and provides justification for the decision in the validator notes column.

3.4.2 Synset gloss validation

The validator provides his validation on the synset gloss column H. The validator can choose between :

- Accepted: If the validator finds that the synset gloss describes the synset correctly and does not contain errors such as spell errors, she/he writes A.

- Rejected: If the validator finds that the synset gloss does not describe the synset or it contains errors, she/he writes R and provides justification.

3.4.3 Synset examples validation

The validator provides his decision on the synset examples in Column K. The validator can choose between:

- Accepted: If the validator finds that the synset examples are correct and they do not contain errors, and if there are no synset examples that may be necessary to describe how to use the lemmas, she/he writes here A. It is possible to accept synsets without examples if the translator did not provide them and the validator accepts the translator decision.

- Rejected: If the validator did not provide examples and the validator does not accept the translator's decision, or if he finds errors in the examples, she/he writes here R and provides justification.

3.4.4 Validator notes validation

The validator provides justification for his rejection of any of the previous synset translations in this column. Validator comments are optional in case of acceptance, but they are mandatory in case of rejection.

In cases where translations were rejected by the validator, the Excel sheet was returned back to the translator with the validator's notes for clarification. The identified mistakes were corrected, or the translators provided reasons for the chosen translations. This process continued until the translators and validators reached an

agreement, and all translations were accepted, making this a high-quality resource.

3.5 Reformatting

The developed resource can be used for the development of NLP applications. However, data in an Excel format is not suitable for programming Integrated Development Environments (IDEs) and computational linguistics fields where these applications are developed (Suárez et al., 2007). This is due to its memory-intensive nature which results in inefficiency. The use of such data in IDEs could lead to increased memory consumption, longer execution time, and reduced performance, making it less than ideal for application development.

To address these limitations, the developed resource was reformatted to Extensible Markup Language (XML) format. XML format is a widely accepted standard for representing structured data (Bourret, 1999). Its standardization ensures consistency and compatibility across different development software applications and platforms. Furthermore, XML provides a machine-readable framework that allows the representation of linguistic data in a hierarchical order and the inclusion of metadata and semantic annotations (Kroeze et al., 2010). The data was grouped according to the parts of speech and converted to XML files.

4. Evaluation and Results

Our validation method explicitly and formally evaluated individual lemma, examples and definitions translations, and their quality. The evaluation was carried out by a group of native Setswana speakers who are proficient in English. They determined the validity of translations by marking them as "accept" if accurate and "reject" if incorrect or lacking in translation equivalents. To calculate the accuracy of the translations, the following metric was used to measure accuracy (A). This is calculated by dividing the number of correct translations i.e. “accept” by the number of total translations using the following equation:

A= (correct translations)/(total number of translations) * 100
A= 5981/6016
A=0.99

There were 5981 synsets correctly translated out of a total of 6016 translated synsets, thus substituting these values into the equation above. The initial submission achieved a 99% accuracy rate before undergoing validation. After validation, continuous revisions and discussions between translators and validators resulted in a unanimous agreement, ultimately achieving a 100% accuracy rate. The presented resource in this study consists of the following translated lexicon in Table 1.

File	Synsets	Number of Words
Setswana-nouns-1	1004	15970
Setswana-nouns-2	1001	15724
Setswana-verbs-1	1001	15643
Setswana-verbs-2	1006	15595
Setswana-adjectives-1	1009	11005
Setswana-adjectives-2	1001	19393
Total number	**6016**	**93330**

Table 1: Translated lexicon statistics.

The number of the lemmas, glosses, and examples are the same for both Setswana and English as all the source data in all rows of all the files were translated.

5. Conclusion

5.1 Bibliographical References

This paper presented a new English-Setswana bilingual dataset that was professionally translated with a specific focus on the space domain. The dataset was developed using the expansion approach, involving the translation of a subset of synsets from PWN into Setswana. The translation process was undertaken by professional Setswana translators specifically contracted for this task. Following the translation process, native Setswana speakers, who also possessed proficiency in English, validated the translated content.

For validation, iterative assessments and discussions between translators and validators confirmed the accuracy of all translations, achieving a 100% accuracy rate. The current funding covered the translation of specified files presented in this paper within this domain, there are still pending files to be translated. Our future endeavours entail securing further funding to significantly enhance the dataset. Additionally, as the current translation process was manual, we aim to semi-automate both translation and validation procedures by leveraging computer-aided translation software. Once completed, the dataset will be openly accessible to researchers for application in linguistic and NLP research.

6. Bibliographical References

Batsuren, K., Ganbold, A., Chagnaa, A. & Giunchiglia, F. Building the mongolian wordnet. Proceedings of the 10th global WordNet conference, 2019. 238-244.

Bella, G., Mcneill, F., Gorman, R., Donnaíle, C. Ó., Macdonald, K., Chandrashekar, Y., Freihat, A. A. & Giunchiglia, F. A major wordnet for a minority language: Scottish gaelic. 12th Language Resources and Evaluation Conference, 2020. European Language Resources Association (ELRA), 2812-2818.

Bosch, S. & Griesel, M. African Wordnet: facilitating language learning in African languages. Proceedings of the 9th Global Wordnet Conference, 2018. 306-313.

Bosch, S. E. & Griesel, M. 2017. Strategies for building wordnets for under-resourced languages: The case of African languages. Literator (Potchefstroom. Online), 38, 1-12.

Bourret, R. 1999. Xml And Databases.

Kroeze, J. H., Bothma, T. J. D. & Matthee, M. C. 2010. Constructing An Xml Database Of Linguistics Data. Td: The Journal For Transdisciplinary Research In Southern Africa, 6, 139-174.

Lastrucci, R., Dzingirai, I., Rajab, J., Madodonga, A., Shingange, M., Njini, D. & Marivate, V. 2023. Preparing The Vuk'uzenzele And Za-Gov-Multilingual South African Multilingual Corpora. Arxiv Preprint Arxiv:2303.03750.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. 2020. Low Resource Language Dataset Creation, Curation And Classification: Setswana And Sepedi. Arxiv Preprint Arxiv:2004.13842.

Mckellar, C. A. & Puttkammer, M. J. 2020. Dataset For Comparable Evaluation Of Machine Translation Between 11 South African Languages. Data In Brief, 29, 105146.

Miller, G. A. 1995. Wordnet: A Lexical Database For English. Communications Of The Acm, 38, 39-41.

Sebolela, F. 2009. The Compilation Of Corpus-Based Setswana Dictionaries. University Of Pretoria.

Suárez, O. S., Riudavets, F. J. C., Figueroa, Z. H. & Cabrera, A. C. G. 2007. Integration Of An Xml Electronic Dictionary With Linguistic Tools For Natural Language Processing. Information Processing & Management, 43, 946-957.

Superman, S., Batman, B., Catwoman, C., and Spiderman, S. (2000). *Superheroes experiences with books*. The Phantom Editors Associates, Gotham City, 20th edition.

7. Language Resource References

Bosch, Sonja and Griesel, Marissa. 2017. Strategies for building wordnets for under-resourced languages: the case of African languages

Compiling a List of Frequently Used Setswana Words for Developing Readability Measures

Johannes Sibeko
Nelson Mandela University
University Way, Summerstrand, Port Elizabeth, 6019, South Africa
johanness@mandela.ac.za

Abstract

This paper addresses the pressing need for improved readability assessment in Setswana through the creation of a list of frequently used words in Setswana. The end goal is to integrate this list into the adaptation of traditional readability measures in Setswana, such as the Dale-Chall index, which relies on frequently used words. Our initial list is developed using corpus-based methods utilising frequency lists obtained from five sets of corpora. It is then refined using manual methods. The analysis section delves into the challenges encountered during the development of the final list, encompassing issues like the inclusion of non-Setswana words, proper names, unexpected terms, and spelling variations. The decision-making process is clarified, highlighting crucial choices such as the retention of contemporary terms and the acceptance of diverse spelling variations. These decisions reflect a nuanced balance between linguistic authenticity and readability. This paper contributes to the discourse on text readability in indigenous Southern African languages. Moreover, it establishes a foundation for tailored literacy initiatives and serves as a starting point for adapting traditional frequency-list-based readability measures to Setswana.

Keywords: Setswana, Frequently used words, Indigenous language, Readability, Low-resourced

1. Introduction

There is consensus that words that are frequently encountered in reading become easier to read (Chen and Meurers, 2016; Rello et al., 2013). This connection between word exposure and ease of reading extends to improved word familiarity and subsequent knowledge (Chen and Meurers, 2016). Conversely, the adverse impact on reading fluency is evident when readers are confronted with unfamiliar words or grammatical structures (Newbold and Gillam, 2010).

Therefore, it becomes imperative to delve into the frequencies of words for the development of readability measures. This awareness of word frequencies can serve as a valuable tool to assess and manipulate levels of text readability.

Understanding text readability is important in South Africa, where literacy levels among language learners are consistently low across various languages, both in home and additional language classes. This concern of low literacy skills is particularly emphasised among language learners such as those in Setswana classes who demonstrate greater proficiency in oral skills than in reading (Lekgoko and Winskel, 2008). According to Mophosho et al. (2019), focusing on enhancing reading proficiency, especially among Setswana learners, is crucial.

While research has been conducted on reading ability in Setswana, such as the work by Pule and Theledi (2023), which delves into challenges in reading proficiency and underscores the influence of prosodic features on Setswana comprehension, and the study by Probert (2019), which advocates for targeted research on reading skills in African languages, pinpointing syllables as crucial units for connected reading in isiXhosa and Setswana, there remains a noticeable lack of knowledge regarding strategies for acquiring reading proficiency in African languages when compared to resource-rich languages like English.

In this paper, we use corpora to develop a list of frequently used words in Setswana. The primary aim of developing this list is to facilitate the adaptation of the Dale-Chall readability index for Setswana.

The rest of this paper provides background to frequency-based readability measures in Section 2, it then discusses the need for measuring text readability in the South African context in Section 3, followed by the method for data collection and analysis in Section 4, the findings that outline problems and solutions in Section 5, a discussion of the findings and the implication of the list of frequently used words in Section 6, and the conclusion with recommendations.

2. Background

Setswana, alternatively referred to as 'Tswana,' 'Chuana,' or 'Sechuana,' is a Bantu language (Bennett et al., 2016). It forms part of the Sotho-Tswana language group with Sesotho and Sepedi. In South Africa, the Sotho-Tswana language group has over 16 million primary speakers (Fraser, 2023).

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 37–44
25 May, 2024.

Setswana constitutes 8.3% (5.15 million speakers), alongside Sepedi (10%, or 6.2 million speakers) and Sesotho (7.8%, or 4.84 million speakers). Although the majority of Setswana speakers are from South Africa, Setswana is also an official language in Botswana, a recognised national language in Zimbabwe, and a marginalised spoken language in Namibia (Otlogetswe, 2001).

Despite the prevalence of Setswana, we are not aware of prior efforts to develop readability measures in the language.

In a review of readability measures, DuBay (2004) notes that over 200 measures have been developed for English, reflecting the extensive scholarship on text readability spanning over two centuries in high-resource languages (Collins-Thompson, 2014; De Clercq and Hoste, 2016; DuBay, 2004).

We adhere to the definition of text readability proposed by Bailin and Grafstein (2001), who define it as the ease with which a text can be read. Our focus on readability does not extend to text comprehension, understandability, extra-textual properties or reader characteristics.

Our search for readability measures for African indigenous languages revealed at least three readability measures for Afrikaans. A comprehensive examination of the three Afrikaans readability measures is presented in (McDermid Heyns, 2007). In essence, the three readability measures for Afrikaans drew inspiration from the English Flesch-Reading Ease measure.

Furthermore, recent developments indicate initiatives to formulate nine readability measures for Sesotho (Sibeko, 2023; Sibeko and Van Zaanen, 2022). These measures encompass four syllable-information-based metrics, four word-length-based metrics, and one frequency list-based metric. For this purpose, Sibeko and De Clercq (2023) crafted a list of frequently used words in Sesotho, intended for incorporation into the development of the Dale-Chall index for Sesotho. Similarly, the efforts in this paper are geared towards the development of a frequency list for inclusion in the Dale-Chall index for Setswana.

Setswana serves various functions including education. Written texts constitute a significant component of communication in Setswana. Consequently, access to written information in Setswana holds paramount importance. Regrettably, despite the inclusion of Setswana in educational curricula from basic to tertiary levels in Southern Africa, a portion of the language users lack formal education, while others possess only limited educational attainment. As a result, the absence of readability measures for these languages poses a significant challenge, especially when readers encounter difficulties extracting information from written communications.

3. Frequency-list-based Measures

A widely accepted hypothesis among readability scholars posits that the readability of a text can be quantified using specific formulas. One prominent category of these formulas includes frequency-list-based readability measures, which operate on the principle that frequently encountered words are easier to recognise, making them easier to read than less common words in texts (Brysbaert et al., 2011). This approach assesses word difficulty by counting infrequently used or challenging words (Gopal et al., 2021). Therefore, the foundation of traditional readability measures, which gauge word familiarity, lies in having a comprehensive list of frequently used words.

To illustrate this principle in practice, George Spache developed the Spache Readability Formula (Spache, 1953). This formula relies on a compilation of familiar words tailored to learners in specific grades. Texts are then segmented into 100-word sections to ascertain the number of unfamiliar words not included in the grade-specific word list (Spache, 1953; Smith, 2016). A higher average of unfamiliar words correlates with harder-to-read texts.

Similarly, Dale and Chall, in their Dale-Chall Index (Dale and Chall, 1948), employ a list of words familiar to and comprehensible by Grade 4 learners. The average of these words is computed, and a higher prevalence of unfamiliar words, absent from the designated list, corresponds to texts that are harder to read.

In this paper, we rely on general corpora and not texts that are tailored for language learners. Even so, our list can serve as a foundation for the development of a frequency-list-based readability measure specifically designed for Setswana.

4. Methodology

We collected five corpora to construct a frequency wordlist, aiming to encompass various genres by gathering texts from different sources. The preparation of each corpus for analysis involved lowercasing using *bash* and tokenisation with *ucto*, including the specific requirement for sentence segmentation. The corresponding sentence information is detailed in Table 1. Below are brief overviews of the five corpora.

4.1. Corpus 1: NCHLT

The objective of the National Centre for Human Language Technology (NCHLT) project was to generate speech and text data to support the development of Human Language Technologies (HLTs) for the 11 official written languages of South Africa

File	Sentences	Lines	Marks	Numbers	Tokens	Words	Types
NCHLT	58 443	58 520	147 058	41 351	1 400 737	1 249 980	38 864
Autshumato	104 976	103 425	266 388	82 062	2 887 117	2 596 847	53 810
PuoBERTa	67 388	67 071	143 214	25 745	2 396 525	2 248 475	41 037
Wikipedia	48 541	47 718	106 459	21 379	1 179 331	1 063 236	42 157
Bible	37 526	30 891	106 386	4 995	958 692	834 748	18 765

Table 1: Summary of Text Properties

(Eiselen and Puttkamer, 2014; Badenhorst and De Wet, 2022). The text collection[1] consists of translated data acquired from the South African Government domain, with ample training and testing samples for language identification tasks in each language (Duvenhage, 2019).

The original dataset includes source texts, lexica, and the corpus (Eiselen and Puttkamer, 2014). We utilised the cleaned corpus data (approximately 1 249 980 words) and not the raw or source files.

4.2. Corpus 2: Autshumato

The Autshumato Machine Translation project[2] developed a translation text corpus for South African indigenous languages. The texts were manually and professionally translated from English into the other ten official written languages of South Africa. The English-Setswana texts are publicly accessible on the South African Centre for Digital Language Resources (SADiLaR) online repository (Mckellar, 2023).

The Autshumato English-Setswana parallel corpora consist of three distinct sets. The Set 1 collection comprises data that has been translated from English into Setswana by professional translators. This set encompasses a total of 324 342 Setswana words. The Set 2 collection contains data sourced as translated file pairs from reliable translators, with a total of 1 099 509 Setswana words. Lastly, the Set 3 collection comprises data crawled from various government websites, containing a total of 1 172 172 Setswana words.

Ultimately, the Autshumato corpus comprised approximately 2 596 847 words. McKellar (2022) outlines at least four text types from the dataset, including magazines, policies, newsletters, and translation works, in addition to documents obtained from the `gov.za` domain.

4.3. Corpus 3: PuoBERTa

We also collected the PuoBerta corpus (Marivate and Wagner, 2023). The PuoBERTa corpus functions as a News Categorisation dataset (Marivate et al., 2023). Its primary objective is to facilitate the development of monolingual resources for Setswana, encompassing tasks such as part-of-speech (POS) tagging, named entity recognition (NER), and mainly, news categorisation. The dataset was derived from online news articles accessible that were provided by the Botswana Government.

The Berta corpus comprises three data files: the development set (230 373 words), the training set (1 806 813 words), and the test set (226 614 words). We amalgamate the texts to compile a corpus of 2 248 475 words pre-processing.

4.4. Corpus 4: Wikipedia

Our Wikipedia corpus is sourced from Leipzig-Corpora-Collection (2020), offering three downloadable corpora. The first corpus, Leipzig-Corpora-Collection (2017), involves texts crawled from general Wikipedia, totalling 660 041 words. The second corpus, from 2018, comprises 232 210 words collected in Botswana. The third corpus, from 2020, consists of 229 987 words from South Africa. Both the 2020 and 2018 files include 10 000 sentences each. In total, our Wikipedia corpus encompasses 1 063 236 words.

4.5. Corpus 5: Bible

We make use of bible texts sourced from the My-Bible project which is a non-profit religious initiative that offers its resources freely at `https://mybible.zone/en/`. This project and website provide Bible translations in various languages, including all the written languages of South Africa. The site provides two Setswana Bible versions including *Beibele e e boitshepo*, the 1907 version that uses the initial and founding orthography of Setswana and *Beibele*, the 1970 version that employs the refined orthography of Setswana. For our paper, we use the 1970 version.

The Bible texts were acquired in SQL3lite format from `https://www.ph4.org/b4_index.php#google_vignette`. All text extraction procedures were executed using *bash* scripts. The Bible texts are categorised into three sections: (i) Bible books with 66 rows of data, (ii) verses with 31,170 rows of data, and (iii) info with 10 rows

[1]Access the NCHLT corpus at `https://repo.sadilar.org/handle/20.500.12185/343`

[2]Access the Autshumato corpus at `https://repo.sadilar.org/handle/20.500.12185/404`

of data. Specifically for our corpus, we extracted verse texts, which then underwent cleaning processes involving the removal of book numbers, chapter titles, and verse information.

Religious texts have been successfully employed for corpus development in previous studies. For instance, Agic and Vulic (2019) utilised parallel articles from the Jehovah's Witness website. Similarly, Marivate et al. (2020) employ Bible texts from both Sepedi and Setswana for a news topic classification task.

4.6. A common frequency list

We generated different frequency lists for the five sets of corpora by calculating frequencies for each. To achieve independence from corpus size, we employed relative frequencies (Brysbaert et al., 2011; Leech et al., 2014; Van Heuven et al., 2014), normalising the frequency lists to occurrences per million tokens. We aimed to extract the most frequent 3 500 words from each set of data. Some of the data sets contained more words on the same level of frequency and thus resulted in longer lists than the intended 3 500 words. The first step resulted in a total of 17 683 words.

Our primary objective was to end up with a list of 3 000 unique words based on the five corpora. To accomplish this, we merged the five lists and ensured the average relative frequencies of duplicate entries. For instance, the entry *'go'* appeared in all five lists with relative frequencies of 28 315.13, 52 336.45, 52 303.89, 61 184.58, and 58 209.73, respectively. The resulting average frequency for this entry is 50 470.156 per million words. We then identified the top 3 000 most frequently used words for the final list.

5. List Analysis

The initial list of 3 000 words was generated automatically using corpus-based frequency measures and later refined through manual processing. The final compilation comprises 3 006 entries, including 2 992 unique entries and 14 instances of varied spellings. The subsequent section provides a detailed account of the curation process involved in finalising the list.

5.1. Non-Setswana Words

We identified a total of 60 instances of non-Setswana words from our initial list. Examples included terms like '*superintendent*,' '*of*,' and '*society*." These instances were excluded due to their lack of Setswana origin and absence of normalised or naturalised Setswana orthography. Nevertheless, contemporary terms such as '*corona*" and '*covid*,' which also deviate from Setswana's naturalised orthography, were retained on the list. This decision was based on the recognition that these terms are more commonly used in the indigenous languages of South Africa than their translated counterparts.

Furthermore, considering linguistic conventions in South Africa, where certain terms like month names, for example, '*June*,' are typically written in English, we have retained these names in the list. However, it is worth noting that not all months are included in the list, as we aim to maintain fidelity to the corpus under analysis. Nonetheless, there are also instances of months in Setswana, such as '*Motsheganong*'.

We also chose to include the entry '*eish*' in our current list. While acknowledging its primary association with a magazine, we opted to retain it due to its additional usage as a borrowed exclamation. This term appears in three of our source corpora, where in the NCHLT corpus, it pertains specifically to the '*Eish*' magazine, and in the Wikipedia and Autshumato corpora, where it is employed both as an exclamation and in reference to the magazine.

5.2. Abbreviations and Acronyms

The initial list included abbreviated words. For example, words such as '*Mopofof*' - representing '*Mopofofesa*' as in professor, '*Moh*' - standing for '*Mohumagadi*' as in Miss, and '*jj*' - for etc. were identified. These abbreviations were retained although full versions for '*Moh*' and '*jj*' were excluded from the list to maintain fidelity to the list.

We also noted that there were instances of unfamiliar abbreviations, such as the ambiguous '*rbn*.' A closer examination revealed that this abbreviation originated from the Autshumato collection, where '*rbn*' referred to a specific company. Consequently, we decided to remove this particular entry from our list.

Secondly, the initial list included acronyms, such as '*SARS*' representing the South African Revenue Service. Note that these entries were anticipated since some texts were sourced from government websites. Among the expected acronyms were '*SAPS*,' denoting the South African Police Service, and '*SASSA*,' an acronym for the South African Social Security Agency, the current distributor of welfare grants in South Africa. Despite this anticipation, we made the decision to eliminate these entries from the list. The rationale behind this choice is twofold. Firstly, these acronyms deviate from the typical Setswana words as they are not normalized into Setswana. Additionally, they demonstrate a specific inclination towards a domain, which further justifies their exclusion.

Even so, we opted to retain globally recognised acronyms such as '*HIV*' (human immune virus),

as they are typical in Setswana texts beyond our current corpus. Interestingly, the Sesotho list of frequently used words (Sibeko and De Clercq, 2023) contains both *HIV* and *aids* while our list only contains *HIV*.

5.3. Proper Names

Our initial list contained at least 80 proper names most of which were biblical names such as '*Gileate, Hesekia, Abesalomo, Jerobeame, Nebukatenesare*' and others. These biblical names were naturalised into Setswana and used expected Setswana orthography.

The list also contained names of African icons such as *Mandela*, as well as names of places such as '*Francistown, Gauteng, Zimbabwe*', and '*Vaal*'. Similar to Sibeko and De Clercq (2023), we removed all instances of proper names. According to Dale and Chall (1948), proper names are automatically deemed familiar and need not be included in the frequency list.

5.4. Multifaceted Meanings

There were instances of words where the meaning was unclear and not immediately discernible without context. For example, the entry '*time*' could be interpreted in English to refer to the passage of time, a specific point in time, or planning. In Sotho-Tswana languages, it can also be used to signify switching off. Similarly, the entry '*rate*' may mean to evaluate or assess in English, but it typically carries the meaning of love in Setswana.

Despite the potential for ambiguity, these words are retained in the final list. This decision acknowledges the diverse meanings they hold across different linguistic contexts. The assumption is that readers will interpret these words in Setswana rather than in English when reading Setswana texts. It is important to note, however, that this ambiguity may pose challenges in the context of multilingual texts where the reader will have to rely on context to aid in identifying the correct language and expected pronunciation when reading.

5.5. Unexpected Words

The preliminary list included unexpected entries. Firstly, there were instances of non-word entries, including numerical values. All such instances were removed from the list as we are interested only in frequently used words.

Secondly, we observed the presence of isolated letters such as '*p*', '*d*', '*g*', '*s*', '*f*', '*i*' and others. We systematically removed all instances of isolated consonants from the list because individual consonants do not qualify as valid Setswana words.

Furthermore, even though certain Setswana vowels can constitute words in a vowel-only context (for instance, '*a,' 'e*' and '*o*'), it was noted that the vowel '*i*' does not serve as a standalone word. Consequently, we excluded this particular entry from our list. Nonetheless, a more thorough analysis revealed that the letter '*i*' was predominantly used as a page number reference in the source documents.

5.6. Spelling and Orthography

There is a general consensus that Sotho-Tswana languages lack specific rules for governing the orthography of loanwords (Chokoe, 2020). This absence of clear regulations manifested in our preliminary list, leading to diverse spellings for the word '*Afrika*.' We identified at least four spelling variations, including '*Africa*,' '*Aferika*,' '*Aforika*,' and '*Afrika*,' with the '*Afrika*' spelling exhibiting a higher frequency. These varied spellings are also associated with related terms such as '*Afrikaborwa*,' '*Afrikan*,' '*Pan-Afrikan*,' '*MoAfrikan*,' '*MaAfrikan*,' and others. Likewise, additional spelling variations yield similar words, as seen with '*MoAforikaborwa*' and '*MaAforikaborwa*,' both present in the list. Like Sibeko and De Clercq (2023), we retained all spelling variations as long as they were part of the initial list of frequently used words.

We were surprised to encounter a misspelled word, namely '*bosetphaba*', which, upon contextual analysis, was identified as originating from the NCHLT corpus. The correct form is '*bosetšhaba*', meaning '*national*'. Recognising it as a typographical error, we have excluded this entry from our current list.

Furthermore, we observed the inclusion of dash-compounded words like '*ba-na-le*' and '*bokone-bophirima*'. To maintain a focus on individual words, we have opted to remove compound entries from our list.

6. Discussion and Conclusion

This paper contributes to the development of readability measures for lower-resourced indigenous languages of Southern Africa by developing a list of frequently used words for Setswana. As detailed in the introduction, the scholarly focus on high-resource languages has left indigenous Southern African languages, including Setswana, understudied in the realm of text readability.

Our research aims to improve the applicability of readability measures to the Sotho-Tswana language group. We draw inspiration from the ongoing Sesotho readability project (Sibeko, 2023), which introduces readability measures based on word length, syllable information, and frequency lists. However, before our work, the transferability

of frequency-list-based measures to Setswana encountered difficulties because there was no curated list specifically tailored for integration into readability measures.

Inspired by established readability measures such as the Dale-Chall Readability Index currently in development for Sesotho, our goal is to adapt and extend these measures to address the unique linguistic context of Setswana. The Dale-Chall Index, known for relying on a list of familiar words, aligns seamlessly with our objective of enhancing readability by prioritising frequently used Setswana words.

6.1. Challenges and Decision-Making

The choice to preserve diverse spelling variations aligns with recommendations in the literature (Sibeko and De Clercq, 2023), highlighting the significance of inclusivity in representing frequently used words. This strategy results in a comprehensive frequency list that acknowledges the linguistic richness and variations present in Setswana. While this approach may introduce a potential mismatch between word use frequency and their inclusion in our list, considering that words may be spelled differently in various corpora, it overlooks the aspect of familiarity for readers encountering different word forms. Consequently, we opted to include words only if they were part of our original shortlist, maintaining fidelity to the actual appearances of words in the list.

Additionally, as illustrated in Section 5, the manual cleaning process revealed non-Setswana words, abbreviations, proper names, and unexpected terms within the Setswana corpus. These occurrences presented challenges during the compilation of the frequency list. Consequently, specific measures were implemented to either retain or exclude these words from the list.

6.2. Implications

The literacy challenges faced by school language learners, particularly those in the Sotho-Tswana language group, underscore the pressing need for tailored readability measures. Unfortunately, Setswana is not well-explored in Natural Language Processing (Marivate et al., 2023). Our goal is to help address this gap by focusing on readability studies specifically in the context of Setswana.

The presence of our list of frequently used words not only offers valuable insights into reading proficiency but also serves as a foundation for the development of Setswana-tailored readability measures that rely on lists of frequently used words. These measures will enable educators and policymakers to make informed decisions, providing targeted strategies to enhance reading proficiency among Setswana learners.

Curriculum developers, assessors, and teachers can leverage our list to guide their language teaching decisions and to select desirable reading materials for both instruction and assessment.

6.3. Limitations

Note that traditional readability measures are criticised for many shortcomings. For instance, according to Crossley et al. (2021), these measures commonly rely on estimates for measuring lexical and syntactic features, while neglecting semantic features and discourse structures, text cohesion and style elements. Furthermore, they are limited in reading criteria and are susceptible to age group and domain specificity. The current paper does not address these shortcomings. Instead, it focuses on the development of a frequency list that can be used in the development of a frequency-based readability measure based on the Dale-Chall index.

The findings presented in this paper exhibit certain limitations. Notably, unlike the Spache Readability Formula examples (Spache, 1953) and the Dale-Chall Index instances (Dale and Chall, 1948), our approach involves compiling a list of frequently used words in the language as observed from limited corpora rather than tailoring it for specific readers in a particular grade level.

While our research was constrained by the absence of originally written texts in Setswana designated for educational purposes and the resulting unavailability of educational corpora, we recommend that future research explores the development of grade-level lists. This refinement could enhance the applicability and precision of readability measures for Setswana, aligning them more closely with the educational context and readership levels targeted in language-related studies.

6.4. Future Directions

Building on our current work, we envision several avenues for future research that will contribute to the ongoing development of Setswana readability measures and broader linguistic studies. Comparative studies with other Sotho-Tswana languages, such as Sepedi and Sesotho, will identify shared linguistic patterns and assess the generalisability of common word lists and readability measures across these languages.

Additionally, extending the analysis of a frequency list such as the one proposed in this paper to include a diverse range of text types, including educational materials, news articles, and literary works, will capture the breadth of Setswana language usage and ensure the applicability of readability measures across contexts. Nonetheless, the

exploration of reading proficiency methodologies in African languages, as advocated by Probert (2019), remains imperative.

7. Bibliographical References

Željko Agic and Ivan Vulic. 2019. Jw300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210. Association for Computational Linguistics.

Jaco Badenhorst and Febe De Wet. 2022. NCHLT auxiliary speech data for ASR technology development in South Africa. *Data in Brief*, 41:107860.

Alan Bailin and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301.

William. G Bennett, Maxine Diemer, Justine Kerford, Tracy Probert, and Tsholofelo Wesi. 2016. Setswana (South African). *Journal of the International Phonetic Association*, 46:235–246.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect. *Experimental psychology*.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th workshop on innovative use of nlp for building educational applications*, pages 84–94.

Sekgaila Chokoe. 2020. Spell it the way you like: The inconsistencies that prevail in the spelling of Northern Sotho loanwords. *South African Journal of African Languages*, 40(1):130–138.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Scott A Crossley, Aron Heintz, Joon Choi, Jordan Batchelor, Mehrnoush Karimi, and Agnes Malatinszky. 2021. The commonlit ease of readability (clear) corpus. In *EDM*.

Edgar Dale and Jeanne Sternlicht Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

William H DuBay. 2004. The principles of readability. Technical report, Impact Information, Costa Mesa.

Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.

Roald Eiselen and Martin Puttkamer. 2014. Developing text resources for ten South African languages. In *Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland*, pages 3698–3703, Paris. European Language Resources Association (ELRA).

Luke Fraser. 2023. These are the most spoken languages in South Africa. Technical report, BusinessTech. Accessed: 11 Jan 2024.

Revathi Gopal, Mahendran Maniam, Noor Alhusna Madzlan, Siti Shuhaida binti Shukor, and Kanmani Neelamegam. 2021. Readability formulas: An analysis into reading index of prose forms. *Studies in English Language and Education*, 8(3):972–985.

Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

Olemme Lekgoko and Heather Winskel. 2008. Learning to read Setswana and English: Cross-language transference of letter knowledge, phonological awareness and word reading skills. *Perspectives in Education*, 26(4):57–73.

Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. 2023. Puoberta: Training and evaluation of a curated language model for setswana. In *Artificial Intelligence Research. SACAIR 2023. Communications in Computer and Information Science*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, Abiodun Modupe, and South Africa CSIR. 2020. Pedi. *arXiv preprint arXiv:2004.13842*.

Jacques McDermid Heyns. 2007. Readability statistics for Afrikaans. In *LSSA/SAALT/SAALA Joint Annual Conference, North-West University, Potchefstroom, South Africa*.

Cindy McKellar. 2022. *Autshumato Monolingual Sesotho Corpus*. ONLINE. South African Centre for Digital Language Resources. Available

at: https://repo.sadilar.org/handle/20.500.12185/583 Accessed: 28 Jan 2023.

Cindy Mckellar. 2023. *Autshumato English-Sesotho Parallel Corpora*. Southern African Centre for Digital Language Resources. Available at: https://repo.sadilar.org/handle/20.500.12185/577[Lastmodified:15Dec.2022].

Munyane Mophosho, Lesedi L Sebole, and Katijah Khoza-Shangase. 2019. The reading comprehension of grade 5 Setswana-speaking learners in rural schools in South Africa: Does home language matter? *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 35(3):59–73.

Neil Newbold and Lee Gillam. 2010. Populating a framework for readability analysis: Word frequency = word difficulty.

Thapelo Otlogetswe. 2001. The BNC design as a model for a setswana language corpus. *extraction*, page 1.

Tracy N Probert. 2019. A comparison of the early reading strategies of isiXhosa and Setswana first language learners. *South African Journal of Childhood Education*, 9(1):1–12.

Violet Maphefo Sefolaro Pule and Kgomotso Theledi. 2023. The impact of the presence of prosodic features (tone markings) on comprehending Setswana words in reading. *African Journal of Inter/Multidisciplinary Studies*, 5(1):1–12.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Human-Computer Interaction–INTERACT 2013: 14th IFIP TC 13 International Conference, Cape Town, South Africa, September 2-6, 2013, Proceedings, Part IV 14*, pages 203–219. Springer.

Johannes Sibeko. 2023. Using classical readability formulas to measure text readability in Sesotho. In Tomaž Erjavec and Maria Eskevich, editors, *Selected papers from the CLARIN Annual Conference 2022*, volume 198, pages 120—132. Linköping Electronic Conference Proceedings, Prague, Czechia.

Johannes Sibeko and Orphée De Clercq. 2023. A corpus-based list of frequently used words in Sesotho. In *Proceedings of the Fourth workshop on Resources for African Indigenous Language (RAIL 2023), Dubrovnik, Croatia*, pages 32–41, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

Johannes Sibeko and Menno Van Zaanen. 2022. Developing a text readability system for Sesotho based on classical readability metrics. In *Digital Humanities Conference: Responding to Asian diversity, Book of Abstracts*, volume 2022, pages 571–572. Short Paper. Available at: https://dh-abstracts.library.virginia.edu/works?keywords=11133 Accessed: 15 Apr. 2023.

Terry Smith. 2016. The problems with current readability methods and formulas: missing that usability design. In *2016 IEEE International Professional Communication Conference (IPCC)*, pages 1–4. IEEE.

George Spache. 1953. A new readability formula for primary-grade reading materials. *The Elementary School Journal*, 53(7):410–413.

Walter JB Van Heuven, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. Subtlex-uk: A new and improved word frequency database for british english. *Quarterly journal of experimental psychology*, 67(6):1176–1190.

8. Language Resource References

Leipzig-Corpora-Collection. 2017. *Tswana community corpus based on material from 2017*. University of Leipzig. PID https://corpora.uni-leipzig.de?corpusId=tsn_community_2017. Accessed: 2024-01-20.

Leipzig-Corpora-Collection. 2020. *Download Corpora Tswana*. University of Leipzig. PID https://wortschatz.uni-leipzig.de/en/download/Tswana#tsn_wikipedia_2021. Accessed: 2024-01-20.

Marivate, Vukosi and Wagner, Valencia. 2023. *Daily News - Dikgang Categorised News Corpus*. Data Science for Social Impact Group. PID https://github.com/dsfsi/PuoBERTa.

A Qualitative Inquiry into the South African Language Identifier's Performance on YouTube Comments

Nkazimlo Ngcungca, Johannes Sibeko, Sharon Rudman
Nelson Mandela University
University Way, Summerstrand, 6019, South Africa
zizingcungca@gmail.com, {johanness, srudman}@mandela.ac.za

Abstract

The South African Language Identifier (SA-LID) has proven to be a valuable tool for data analysis in the multilingual context of South Africa, particularly in governmental texts. However, its suitability for broader projects has yet to be determined. This paper aims to assess the performance of the SA-LID in identifying isiXhosa in YouTube comments as part of the methodology for research on the expression of cultural identity through linguistic strategies. We curated a selection of 10 videos which focused on the isiXhosa culture in terms of theatre, poetry, language learning, culture, or music. The videos were predominantly in English as were most of the comments but the latter were interspersed with elements of isiXhosa, identifying the commentators as speakers of isiXhosa. The SA-LID was used to identify all instances of the use of isiXhosa to facilitate the analysis of the relevant items. Following the application of the SA-LID to this data, a manual evaluation was conducted to gauge the effectiveness of this tool in selecting all isiXhosa items. Our findings reveal significant limitations in the use of the SA-LID, encompassing the oversight of unconventional spellings in indigenous languages and misclassification of closely related languages within the Nguni group. Although proficient in identifying the use of Nguni languages, differentiating within this language group proved challenging for the SA-LID. These results underscore the necessity for manual checks to complement the use of the SA-LID when other Nguni languages may be present in the comment texts.

Keywords: Language Identity, IsiXhosa, Language Identification, SA-LID

1. Introduction

The global linguistic landscape, comprising approximately 7168 languages, is dynamic and demands continuous exploration (Aitchison, 2005; Trask, 2003). This is particularly true in light of the core role played by language in terms of the social, cultural, intellectual and political vitality in any society (Lo Bianco, 2010). As such, there is a need for continuing research in order to understand the characteristics of each language as well as the cultures and identities that are linked to the concerned linguistic communities.

Given the global linguistic diversity, an ability to distinguish between the languages being used in a particular context is understandably significant (Jaspers, 2015). Such an ability facilitates the decoding of the content of the message and thus fosters effective communication and comprehension (Hardan, 2013). This is particularly significant in a linguistically diverse country like South Africa (Fishman et al., 2008), where most citizens are multilingual (Evans, 2015; Sithole, 2015; Adelani et al., 2021).

Through language, individuals not only communicate but also articulate their origins, making it a fundamental dimension of cultural identity. In this, and many other ways, language and identity are intricately linked (Bucholtz and Hall, 2004). This aspect of language use extends beyond oral expression to also encompass written interactions and specifically so in colloquial contexts which allow for a more spontaneous and free use of language - for example, social media platforms. A tool which has the ability to accurately identify the languages used in a multilingual text could carry numerous benefits and play an important role in a linguistically diverse society. This would be especially true in terms of research focused on the actual use of language by those fluent in more than one language and the manner in which their language use expresses their cultural identity.

Identity, in its simplest form, is an expression of individuality and reflects the uniqueness of every human being (Buckingham, 2008). However, identity is also influenced to a large extent by the social groups to which an individual belongs (Baxter, 2016). This is particularly so in terms of the cultural and linguistic background into which one is born as this is the context in which one first learns about – and learns how to express – aspects of the world (Praeg, 2014). Linguistic and cultural identity are generally conflated and language use is often reflective of these aspects – along with other 'hints' about a speaker's identity (Bucholtz and Hall, 2005). For this reason, the relationship between language usage and cultural/linguistic identity is rife with possibilities.

Research on the link between language use and cultural or linguistic identity in a multilingual con-

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 45–54

text assumes the ability to discern between the languages employed by the language users in that community. This is indeed the case with the research project of which this paper forms a component. The broader study aims to investigate linguistic strategies employed by isiXhosa language users to express their language identities on YouTube through the use of comments.

The scope of this study, therefore, underscores the necessity of a reliable language identifier to accurately detect the language(s) used within a particular text. The use of such a tool becomes indispensable when navigating through the substantial pool of comments in order to extract comments written in isiXhosa or code-switched between isiXhosa and other languages. The identification of instances of isiXhosa usage from our corpus of YouTube comments is thus necessary in order to delineate the data on which our study will focus.

This paper aims to evaluate the reliability of the South African Language Identifier (Puttkammer et al., 2016) when it is applied to a corpus of YouTube comments to ascertain the languages used. The following sections of this paper provide a brief literature review in terms of the core concepts in Section 2, an overview of our methodology for data collection in Section 3 and analysis as well as a summary of our findings in Section 4. The paper concludes with a discussion of our conclusions and recommendations in Section 5.

2. Background

2.1. The isiXhosa Language

While there are between 24 and 30 spoken languages in South Africa (Finlayson and Madiba, 2002), the constitution of the Republic of South Africa recognises 12 of these as official languages (Republic of South Africa, 1996, 2023). These official languages are typically grouped into six language groups, including: (i) South African Sign Language, (ii) Sotho-Tswana, which includes Sesotho, Setswana, and Sepedi, (iii) Sotho-Makua-Venda, which includes Tshivenḓa, (iv) West Germanic, which includes Afrikaans̆ and English, (v) Nguni-Tsonga, which includes Xitsonga, and (vi) Nguni, which includes isiZulu, Siswati, siNdebele and isiXhosa.

The Nguni language group occupies a significant position as the largest language group in South Africa. IsiXhosa, the second-most prominent Nguni language within South Africa, (Wheeler, 2018), is predominantly spoken in the Eastern Cape and the Western Cape Provinces. Notably, it is also officially recognised in Zimbabwe (Republic of Zimbabwe, 2021). According to Wheeler (2018), isiXhosa has much in common with isiZulu in terms of their linguistic roots. In fact, as discussed later in Section 4.3, isiXhosa demonstrates close linguistic ties and mutual intelligibility with other languages in the Nguni group as well (Dyers, 2000). Additionally, isiXhosa stands out for its use of clicks, a feature present in only about 0.5% of the world's languages (Brenzinger and Shah, 2023), including a few Bantu languages. These clicks are represented by the use of three consonants: /c, q, and x/ (Nogwina et al., 2013; Gxowa-Dlayedwa, 2015, 2018; Wheeler, 2018).

2.2. Identifying Languages

The initial step in comprehending written text is to ascertain the language in which it is written (Babhulgaonkar and Sonavane, 2020). Various language identification tools are developed for this purpose, with the goal of discerning the language(s) present in the text (Jauhiainen et al., 2024). Note that these language identifiers are designed to encompass both speech and written texts. However, due to the inherent differences between written text, composed of discrete characters, and speech, which involves a continuous signal relying on acoustic features, different natural language processing methods are traditionally employed for text and speech, resulting in limited methodological overlap between the two (Murthy and Kumar, 2006; Ambikairajah et al., 2011).

For the purpose of this discussion, our focus is specifically on language identifiers for written texts. Text language identification involves analyzing written linguistic features, including character n-grams and word frequency patterns. This analytical process often makes use of statistical models and machine learning algorithms (Nezhadi et al., 2017).

Traditionally, human beings are regarded as the most accurate language identifiers (Deshwal et al., 2020). Unfortunately, their ability to detect languages is limited by their language repertoires. As such, the limits of relying on humans for language identification become obvious when considering the estimated 7168 languages worldwide (Al-Jarf et al., 2022), or the twelve official languages in South Africa. Simply put, humans are unable to detect languages that are outside their current linguistic repertoires.

As a result, more non-human dependent approaches are needed in the task of identifying languages. Over time, computational approaches employing tailored algorithms and indexing structures have developed to discern language usage without human intervention (Calvo et al., 2017). This evolution includes the use of advanced techniques such as neural networks (Talpur and O'Sullivan, 2020) and Natural Language Processing (NLP) approaches (Saji et al., 2022), which are integrated into language identification tools.

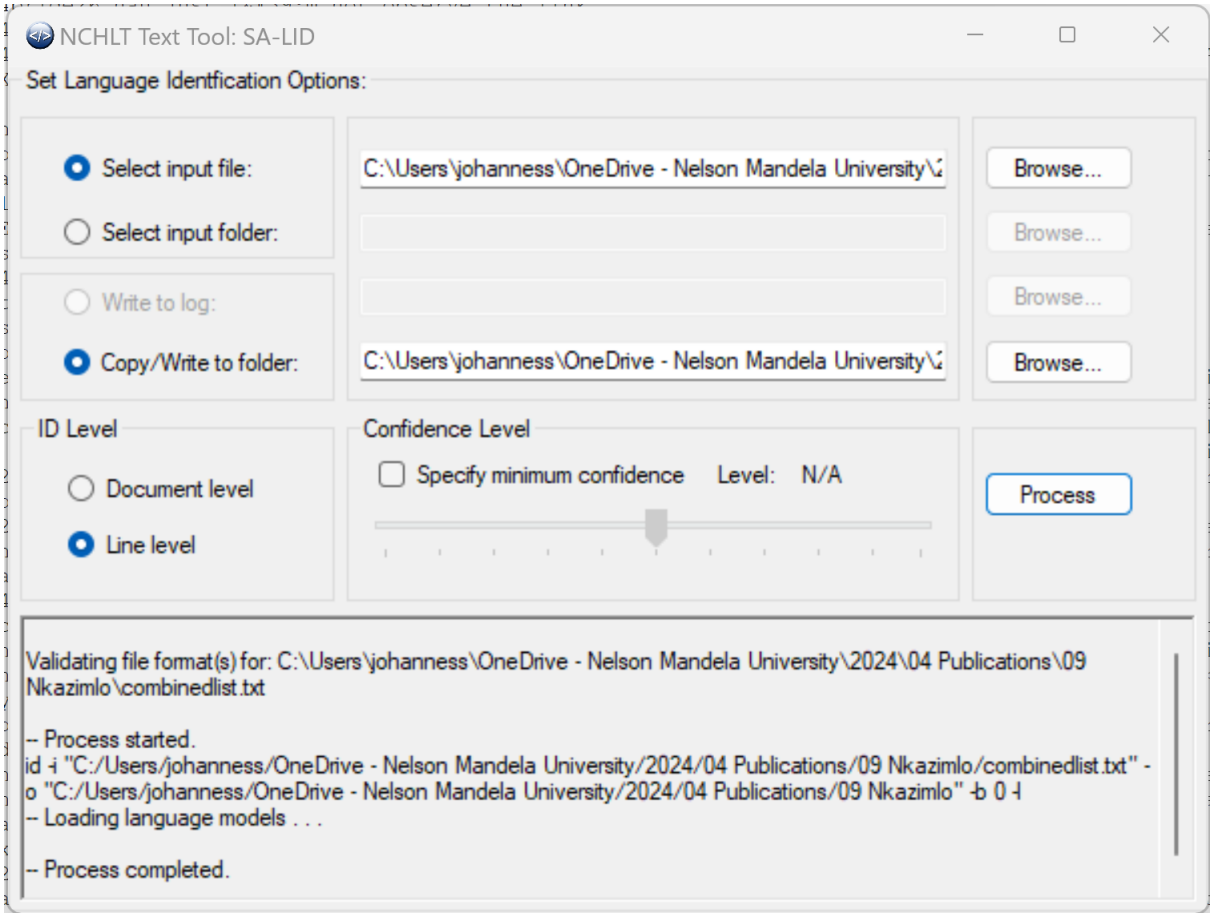


Figure 1: A screenshot of the SA-LID.

2.3. Language Identification Tools

The introduction of automatic language identifiers serves as a valuable advancement in language detection. The primary task of a language identification tool is to analyse a given spoken or written text and generate a prediction of the language in which the text is spoken or written (Navrátil, 2006; Bergsma et al., 2012; Solorio et al., 2014). This process includes assessing the probability of each word in the provided text as belonging to one or more of the languages in the tool's library (Lui and Baldwin, 2012). The identified language is determined by the highest probability, initiating a competition among language models to determine the most likely match for the entire sample. Like humans, automatic language identifiers also rely on libraries (Agarwal et al., 2023). In this way, the automatic tools need to be trained using different languages and will not be able to detect new languages that are not in their existing libraries.

In this paper, as indicated in Section 1, we conduct a qualitative evaluation of the South African Language Identifier (henceforth SA-LID). The SA-LID was designed to classify text into one of the 11 official written languages of South Africa, either at the document or line level (Puttkammer et al., 2018). The SA-LID has been trained using government text corpora obtained during the National Centre for Human Language Technology (NCHLT) Text project and collected through collaboration between the South African Department of Arts and Culture and the Centre for Text Technology (Puttkammer et al., 2018).

The SA-LID uses feature extraction, identifying language-specific patterns through the analysis of character n-grams, ranging from bigrams to 6-grams. The model was trained using the Multinomial Naive Bayes classifier, incorporating labelled training samples and the selected feature extractor. These components collectively enable the model to discern languages effectively based on the extracted features. In the subsequent step of text classification, the trained classifier is applied to text inputs, resulting in a list of probable languages arranged by their respective probabilities. The final language determination is achieved by selecting the language with the highest probability, based on the model's consideration of learned patterns and characteristics during the training process, ensuring accurate identification of the language in the given text.

Video Identifier	Release Year	Comments	Views-to-Date	Likes-to-Date	Type
iZcx_akfXe4	2010	202	306,211	870k	Documentary
zEoYl4Ok6Ks	2012	107	394,025	2k	Music and Dance
baEiWB2aM9Y	2013	894	1,797,004	17k	Interview
ZnnlJzINWs8	2018	179	426,724	4.9k	Praise Poetry
ZcRykTbiva4	2015	236	373,909	4.7k	Lesson
zOUvWM6Yx3Q	2015	115	521,804	1.7k	Drama
RfcnDHYFETs	2020	266	14,458	345k	Lesson
rjo8h5qLpU0	2020	2549	2,230,543	92k	Music
v4iOTPFz0-c	2021	1538	222,000	3.5k	Documentary
zPM8Qid9VSY	2021	831	121,577	4.6k	Lesson
Total	-	**6885**	**6,127,486**	**126.7k**	

Table 1: Video Statistics (Ordered by Release Year) with Totals

2.4. Related Research

Previous research has explored the application of language identification specifically to isiXhosa texts (Kyeyune, 2015). In their study, Kyeyune (2015) utilised corpora from the Language Resource Management Agency and employed the Java Text Categorising Library to extract *n*-grams for identifying isiXhosa using an *n*-gram language model. The study conducted by Duvenhage et al. (2017) investigated the use of a naive Bayes classifier for accurate language group identification. Additionally, they incorporated a lexicon-based classifier to differentiate the specific South African language in which the text is composed. Furthermore, in their work, Duvenhage (2019) introduces a hierarchical classifier that combines naive Bayesian and lexicon-based approaches for short-text Language Identification (LID). This approach proves particularly beneficial for under-resourced languages.

In this paper, we investigate the reliability of the SA-LID to assess its usability in detecting isiXhosa from YouTube comments, employing a qualitative approach for our discussion.

3. Methodology

A total of ten videos were selected from YouTube using a variety of pre-determined search terms such as: (i) amaXhosa ase South Africa, (ii) Introduction to the Xhosa culture (iii) The History of isiXhosa language, (iv) The history of isiXhosa culture, (v) Clicks used in isiXhosa music, and (vi) isiXhosa language-use in South Africa. The video selection process was based on the relevance to the title of the broader study, as well as evidence of the use of linguistic elements which identified commentators as isiXhosa.

We employed the YouTube API to mine comments from the 10 selected videos for our study. This process involved specifying the video IDs of the chosen content and extracting the associated comments. The API facilitated the extraction of text-based comments and emojis. We excluded information such as user details, timestamps, and other information.

The data collection was conducted on 19 January 2024. We identified videos that were uploaded more than a year before our investigation. As a result, we were not expecting any surge in the new comments on the videos.

3.1. Data Cleaning

During the data cleaning process, we addressed the presence of unexpected characters by replacing them with relevant punctuation. For example, we transformed $(\&\#39;)$ into the apostrophe $(')$, resulting in modifications for a total of 2371 + 83 instances. Additionally, occurrences of $(\")$ were replaced with $(“)$ and $(”)$, with a total of 730 errors identified and rectified. Furthermore, instances of $(\<3)$ were amended to <3, with twelve occurrences addressed. Finally, we removed line breaks that were indicated by $(< br >)$ as we needed the comments to be counted as one and not to be separated.

This study employed the YouTube Data API to systematically extract comments and replies from specified YouTube videos using associated video IDs. The video IDs are provided in Table 1. An API key was configured for authentication, and a systematic approach was adopted to retrieve comments and replies for each video. The script iterated through the list of video IDs, employing the YouTube Data API to retrieve comments in batches of 100, with pagination support for handling larger datasets. The retrieved comments and replies were processed and organised into a pandas DataFrame for each video, facilitating subsequent analysis.

The dataset initially comprised 6885 lines. However, this figure was reduced after we eliminated duplicate lines. punctuation-only lines (such as question marks), and closing quotation marks (indi-

Language	Confidence levels						
	40%	**50%**	60%	70%	80%	90%	99%
Afrikaans	13	**75**	4	1	1	1	-
English	4225	**5215**	2153	896	177	24	-
SiNdebele	14	**37**	6	3	1	1	-
Sepedi	2	**13**	1	1	1	1	-
SiSwati	12	**44**	3	1	1	1	-
Sesotho	3	**20**	2	-	-	-	-
Setswana	4	**15**	-	-	-	-	-
Xitsonga	2	**20**	-	-	-	-	-
TshivVenda	2	**19**	1	-	1	-	-
IsiXhosa	340	**458**	148	75	27	5	-
IsiZulu	194	**318**	85	48	13	3	-
Unsure	1762	**339**	4170	5548	6352	6539	6573

Table 2: Results considering different confidence levels.

cating the end of quotes from preceding lines) and instances involving full stops, numbers and further duplicates. Ultimately, our analysis is grounded in a dataset encompassing 6573 lines, inclusive of both text and emoji comments. The overview of videos and comment counts is provided in Table 1.

The study prioritised user privacy and adhered to the terms of service of the YouTube platform. No personally identifiable information was collected, and the data were used exclusively for research purposes.

3.2. Data Analysis

We considered seven confidence levels available through the Language Identifier (namely 40%, 50%, 60%, 70%, 80%, 90%, and 99%) in order to evaluate the consistency of the findings. In the end, our discussion is based on the results of the default confidence setting for language identification, that is, 50% confidence. The results of the analyses at the different confidence levels are presented in Table 2.

The SA-LID utilises an input file or folder, and the accepted file types are .txt files. It then outputs to either a log file or into a folder. The identity levels include document level and line level. It is important to note that when the line level option is selected, the output setting defaults to the "Copy/Write to folder" option. A screenshot of the interface is illustrated in Figure 1. The output files classify the sentences, so each output file includes only sentences identified as the specified language. The output file names append the language code as a prefix to the original file name. For example, when using the original file name for a bilingual dataset for English and isiZulu, dataset.txt, the SA-LID will output zu.dataset.txt and en.dataset.txt.

4. Qualitative Error Analysis

In evaluating the performance of the SA-LID on our YouTube corpus (identified for our larger project on the language identities of amaXhosa), we employed a qualitative error analysis. Our analysis was based on the default confidence setting, specifically the 50% confidence level (refer to Table 2 for the results of the SA-LID which reflects the different confidence levels from 40% to 99%).

4.1. The Unsure Caterogy

The SA-LID encountered 339 comments which it found uncertain. Upon examination, we identified a few reasons for the uncertainty. The uncertainty arose primarily from the identification of emojis and unfamiliar slang, such as *'wow'* and *'yeah'* as well as acronyms such as *'lol'* and *'omg'*. We assume that such words were not included in the development data for the SA-LID.

Secondly, a notable challenge emerged as we realised that the SA-LID encountered difficulties in accurately categorising language when spelling mistakes were present. Consequently, a significant number of comments ended up in the unsure category. For instance, one comment under the unsure category featured the misspelling *'qween*,' which is appropriately spelt as *'queen*.' These instances illustrate that when words are misspelt, it becomes more challenging for the SA-LID to accurately identify the languages. This underscores the critical importance of accurate spelling for the SA-LID to perform effectively in language categorisation.

Thirdly, notable instances of unexpected scripts were observed. For instance, the data contained comments in Japanese, Russian, and Arabic. Such scripts are not official in South Africa and, as such, they are not expected to be identified using the SA-LID.

We also noted that some comments in English were also categorised as unsure. For example, consider the comment:

```
(a) Nice, Lucky you, I am so Jealous.
```

We are unable to account for these results. However, such occurrences prompt questions about whether the majority of words in such comments were absent from the language library used by the identifier.

4.2. Multilingual Comments

Ideally, the SA-LID as outlined earlier, will assign the language based on the higher probability. As an example, consider the sentence below:

```
(b) Awume kancane wena. uShaka
uhlanganaphi nokubaleka kwamaXhosa.
Ehamba nabelungu. Babuya Kuphi?
Asibafuni iningi lethu Kwazulu
KwaZulu. Loyalt to nothing
asibafuni.
```

The SA-LID has categorised this sentence under isiZulu since it contains 15 isiZulu words although it also contains three English words, one of which is spelled incorrectly.

Furthermore, we observed that the SA-LID shows a preference for indigenous South African languages when an equal number of words from multiple languages are present in the same comment. To illustrate, consider the following comment classified under the isiXhosa comments:

```
(c) This woman is talking sh\%t...
lo othi xa ungenamgidi awuyndoda...
mxfm.... The only part I like is
lena athi yimisebenzi yakho
ebonakalisa ubudoda.
```

In this example, the term *'sh%t'* is not recognised as English due to the inclusion of punctuation. Upon tokenisation, the word is divided into three tokens, making it less likely to be identified as a valid English word. Additionally, the term *'mxfm'* is a misspelt word. Consequently, there are only ten English words in the comment. Similarly, the isiXhosa word *'awuyndoda*' is spelt incorrectly, as such there are also ten isiXhosa words. Thus, the comment contains an equal number of English and isiXhosa words but even so, the SA-LID identified the comment as isiXhosa, thereby illustrating a preference for isiXhosa.

In more extreme circumstances, the SA-LID identified code-switched comments that are predominantly English under indigenous languages such as isiXhosa. To illustrate, consider the example below:

```
(d) Singabantu abanye.
Xhosas from Zim moved to Zim
from the Eastern Cape with Cecil
John Rhodes, Ndicelumenywa.
```

This code-switched comment exhibits a prevailing use of English, interspersed with three isiXhosa words. The classification of this sentence as isiXhosa reinforces our inference that the SA-LID tends to favour indigenous languages when categorising code-switched texts.

Note that the SA-LID has no category for sentences that are multilingual. This is particularly problematic in the context of South African multilingual social media. That is, there is a need for an additional category in terms of those sentences considered 'multilingual' (rather than assigning them to one of the two language groups present in the sentence). The ability to identify the use of more than one language within a single text effectively ensures an alignment with real-life language use. However, the SA-LID currently identifies at least one language from the comment and then assigns a language label rather than noting the sentence as 'multilingual. Nonetheless, this ability to classify code-switched texts is a significant asset for our study which focuses on how amaXhosa articulate their linguistic identity.

Our aim in this paper was to investigate the ability of the SA-LID to identify comments in isiXhosa. Overall, the SA-LID was able to identify instances of the use of isiXhosa including sentences that are purely in isiXhosa and those that are code-switched.

In the larger study on language identities, we also hope to identify and analyse strategies used by multilingual commentators in their interactions on YouTube as a social media platform. Consequently, the accurate identification of isiXhosa through the SA-LID holds particular significance for our research objectives, facilitating the exclusion of comments lacking isiXhosa content.

4.3. Mutual Intelligibility

In our analysis, we observed challenges for the SA-LID in distinguishing between similar languages from the same language group. For instance,isiZulu and isiXhosa share some characteristics, enabling speakers of one language to understand the other due to their akin dialects.

While there are some similarities in vocabulary stemming from their common Bantu origin, specific words differ between isiXhosa and isiZulu. The table below provides an illustrative example:

Despite these distinctions, the SA-LID encountered difficulty and misidentified some texts written in isiZulu as isiXhosa. For instance, consider the following example:

English	isiXhosa	isiZulu
I want (it)	Ndiyayifuna	Ngiyayifuna
I noticed that/it	Ndiyibonile	Ngiyibonile
I am happy	Ndiyavuya	Ngiyajabula
We appreciate	Siyakuvuyela	Siyakujabulela

(e) Ngiyalithanda isiko lamaXhosa, thanks for this content bhudi''

In this example, the term *'ngiyalithanda*' is of isiZulu origin, while the isiXhosa equivalent would be *'ndiyalithanda'*. We suspect that the confusion might have arisen due to the inclusion of the term *'lamaXhosa*' in the sentence. Nevertheless, the term *'isiko*' can be identified in either of the two languages. Furthermore, examples such as:

(f) ''nazoke ezakuthi madoda''
(g) ''gaaa ! hlala phansi.''

The first example was identified as isiNdebele, while the second example was identified as Siswati. While these may be correct, the same sentences could be identified as other languages in the Nguni group too. For instance, while the use of the word *'ezakuthi'* in the first example rules out isiZulu, which would be *'ezakithi'*, it can be identified as isiXhosa. However, the second example could be isiZulu because of the word *'phansi'*, whose equivalent in isiXhosa is spelled *'phantsi'*. Note that local dialects may actually identify these languages as either one in the group based on language contact influences. Nonetheless, these examples demonstrate the mutual intelligibility of the languages. Furthermore, this illustrates that a thorough manual check is necessary to distinguish between the Nguni languages before commencing an official analysis, as the SA-LID may be confounded by the linguistic similarities.

4.4. Assumed linguistic and Cultural identities

As we analyzed the comments, we observed a diverse array of languages employed by commentators, including code-switching between indigenous South African languages as well as the unexpected occurrences of Japanese, Russian, and Arabic. We inferred that individuals who use both monolingual and multilingual sentences were concurrently expressing both their thoughts and cultural identities. Drawing on the insights of scholars like Bucholtz and Hall (2004) and others who explore the intricate relationship between language and identity, our findings suggested that commentators were strategically situating their linguistic and cultural identities through their language use.

Nevertheless, we acknowledge the inherently multilingual nature of the world, where individuals can learn languages beyond those spoken at home. According to Kinginger (2004), when individuals speak and learn a new language, they simultaneously adopt a new identity or engage in the reconstruction of their existing one. This concept is illustrated by Johanson Botha (2009)'s example of an English man learning isiXhosa. When he speaks isiXhosa, he becomes loud, which causes embarrassment to his wife. This loudness, not commonly associated with Western culture, is stereotypically linked to isiXhosa culture, portraying the construction of amaXhosa as assertive or loud. Therefore, we understand that to definitively ascertain whether someone identifies as isiXhosa or any other language, further investigation (for example conducting interviews) would be imperative.

Given this context, the primary investigation of our broader study will focus on conducting interviews to confirm linguistic identities. This was not, however, necessary for this paper as the objective was solely to assess the accuracy of the SA-LID when applied to a corpus of YouTube comments.

5. Conclusion

In this paper, we explore the use of automatic language identification using the South African Language Identifier (SA-LID) to discern languages within a YouTube comments corpus. This study forms part of a broader project aiming to uncover linguistic strategies employed by isiXhosa speakers in expressing their language and cultural identities in YouTube comments. As digital platforms continue to shape communication patterns, understanding language identities becomes crucial for fostering inclusive and accurate representation. To facilitate this, there is a need for accurate language identification in multilingual texts. As such, the context of our broader study led us to evaluate the reliability of the SA-LID in identifying any use of isiXhosa language elements in the relevant comments which we mined from YouTube. The question which underpinned our research, as reflected in Section 1, related to whether we could rely on the language identification results generated through the use of the SA-LID to accurately identify all instances of the use of isiXhosa.

Our analysis of the SA-LID reveals both strengths and challenges. The tool demonstrates proficiency in identifying languages used in multilingual comments, showcasing its versatility in capturing dynamic language use within the amaXhosa community on YouTube. This aspect prompts us to conclude that the SA-LID can indeed be effectively employed in situations where two languages coexist or code-switching occurs, showcasing its robust capabilities in language categorisation.

However, challenges arise, particularly in cases

of mutual intelligibility between closely related languages like isiXhosa and other Nguni languages. This highlights the complicated nature of language identification and, therefore, urges further exploration.

Other challenges encountered with this tool include uncertainties related to emojis, slang, unconventional spelling and spelling errors. This emphasises the need for continuous refinement in language identification tools to accommodate diverse linguistic expressions. In the context of our corpus, the non-Latin scripts in the dataset further complicated language identification as they are unexpected in the South African context.

Our findings contribute to the ongoing discourse on language use and identity in digital spaces, offering insights into methodologies which can be employed in further research. The misidentification of languages, as noted in this study, opens up opportunities for future studies to explore how the choice of words or phrase structure in a text can potentially confuse a language identifier. In this study, we did not delve into grammatical complexities or sentence structures; our primary focus was to ascertain the ability of the SA-LID to identify the use of isiXhosa from written YouTube comments accurately. We acknowledge that potential issues may have arisen from variations in pre-processing steps. Specifically, different processes may have been employed for tokenisation, text normalisation, and handling special characters compared to those used in the training of the SA-LID.

6. Bibliographical References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages. *arXiv preprint arXiv:2210.11744*.

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

DOSUL ÁFRICA. 2020. Constitution of the republic of south africa, 1996. *As adopted*, 704:705.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. Limit: Language identification, misidentification, and translation using hierarchical models in 350+ languages. *arXiv preprint arXiv:2305.14263*.

Jean Aitchison. 2005. Language change. In *The Routledge Companion to Semiotics and Linguistics*, pages 111–120. Routledge.

Reima Al-Jarf et al. 2022. Text-to-speech software for promoting efl freshman students' decoding skills and pronunciation accuracy. *Journal of Computer Science and Technology Studies*, 4(2):19–30.

Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu. 2011. Language identification: A tutorial. *IEEE Circuits and Systems Magazine*, 11(2):82–108.

Arun Babhulgaonkar and Shefali Sonavane. 2020. Language identification for multilingual machine translation. In *2020 International Conference on Communication and Signal Processing (ICCSP)*, pages 401–405. IEEE.

Judith Baxter. 2016. Positioning language and identity: Poststructuralist perspectives. *The Routledge handbook of language and identity*, pages 34–49.

Kenneth R Beesley. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th annual conference of the American Translators Association*, volume 47, page 54.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific twitter collections. In *Proceedings of the second workshop on language in social media*, pages 65–74.

Matthias Brenzinger and Sheena Shah. 2023. A typology of the use of clicks. *Stellenbosch Papers in Linguistics Plus*, 67(1):59–77.

Mary Bucholtz and Kira Hall. 2004. Language and identity. *A companion to linguistic anthropology*, 1:369–394.

Mary Bucholtz and Kira Hall. 2005. Identity and interaction: A sociocultural linguistic approach. *Discourse studies*, 7(4-5):585–614.

David Buckingham. 2008. *Introducing identity*. MacArthur Foundation Digital Media and Learning Initiative.

Rafael A Calvo, David N Milne, M Sazzad Hussain, and Helen Christensen. 2017. Natural language processing in mental health applications using non-clinical texts. *Natural Language Engineering*, 23(5):649–685.

Deepti Deshwal, Pardeep Sangwan, and Divya Kumar. 2020. A language identification system using hybrid features and back-propagation neural network. *Applied Acoustics*, 164:107289.

Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. *arXiv preprint arXiv:1911.07555*.

Bernardt Duvenhage, Mfundo Ntini, and Phala Ramonyai. 2017. Improved text language identification for the south african languages. In *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, pages 214–218.

Charlyn Dyers. 2000. *Language, identity and nationhood: Language use and attitudes among Xhosa students at the University of the Western Cape, South Africa*. Ph.D. thesis, University of the Western Cape.

John Edwards. 2009. *Language and identity: An introduction*. Cambridge University Press, New York.

Moyra Sweetnam Evans. 2015. Language use and language attitudes in multilingual and multicultural south africa. *Heritage and exchanges: Multilingual and intercultural approaches in training context*, pages 43–62.

Rosalie Finlayson and Mbulungeni Madiba. 2002. The intellectualisation of the indigenous languages of South Africa: Challenges and prospects. *Current issues in language planning*, 3(1):40–61.

Joshua A Fishman, Monica Barni, and Guus Extra. 2008. *Mapping linguistic diversity in multicultural contexts*. Mouton de Gruyter.

Ntombizodwa Gxowa-Dlayedwa. 2018. Investigating click clusters in isixhosa syllables. *South African Journal of African Languages*, 38(3):317–325.

Ntombizodwa Cynthia Gxowa-Dlayedwa. 2015. Ukufundisa izicuku zeziqhakancu emagameni. *Per Linguam: a Journal of Language Learning= Per Linguam: Tydskrif vir Taalaanleer*, 31(3):32–48.

Abdalmaujod A Hardan. 2013. Language learning strategies: A general overview. *Procedia-Social and Behavioral Sciences*, 106:1712–1726.

Jürgen Jaspers. 2015. Modelling linguistic diversity at school: the excluding impact of inclusive multilingualism. *Language Policy*, 14:109–129.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister LindÚn. 2024. Introduction to language identification. In *Automatic Language Identification in Texts*, pages 1–17. Springer.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

Liz Johanson Botha. 2009. 'them and us': Constructions of identity in the life history of a trilingual white south african. *African Identities*, 7(4):463–476.

Celeste Kinginger. 2004. Alice doesn't live here anymore: Foreign language learning and identity reconstruction. *Negotiation of identities in multilingual contexts*, 21(2):219–242.

Michael J Kyeyune. 2015. Isixhosa search engine development report. Technical report, University of Cape Town.

Joseph Lo Bianco. 2010. The importance of language policies and multilingualism for cultural diversity. *International Social Science Journal*, 61(199):37–67.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proceedings of the 5th workshop on language analysis for social media (LASM)*, pages 17–25.

Thembinkosi Mtonjeni. 2013. *An investigation of discriminatory language used in communicating with South Africans born in Tanzania and Zambia*. Ph.D. thesis, Stellenbosch: Stellenbosch University.

Kavi Narayana Murthy and G Bharadwaja Kumar. 2006. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(01):57–80.

Jirı Navrátil. 2006. Automatic language identification. *Multilingual speech processing*, pages 233–272.

Mohammad M Alyan Nezhadi, Majid Forghani, and Hamid Hassanpour. 2017. Text language identification using signal processing techniques. In *2017 3rd Iranian Conference on Intelligent Systems and Signal Processing (ICSPIS)*, pages 147–151. IEEE.

Mnoneleli Nogwina, Zelalem Shibeshi, and Zoliswa Mali. 2013. Towards developing a stemmer for the isixhosa. In *WIP. SATNAC Conference*.

Bonny Norton. 2010. Language and identity. *Sociolinguistics and language education*, 23(3):349–369.

Jacobus Christiaan Oosthuysen. 2016. *The grammar of isiXhosa*. African Sun Media.

Leonhard Praeg. 2014. *A report on Ubuntu*. University of KwaZulu-Natal Press Pietermaritzburg.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. Nlp web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49.

Republic of South Africa. 1996. *Constitution of the Republic of South Africa*. Department of Justice, Pretoria.

Republic of South Africa. 2023. *Constitution Eighteenth Amendment Bill*. Department of Justice and Correctional Services, Pretoria.

Republic of Zimbabwe. 2021. *The Constitution of Zimbabwe*. Veritas, Harare.

Lourdes C Rovira. 2008. The relationship between language and identity. the use of the home language as a human right of the immigrant. *REMHU-Revista Interdisciplinar da Mobilidade Humana*, 16(31):63–81.

Ami Katherine Saji, Laura Morales, and Meredith Winn. 2022. *Feasibility report on setting up a collection on questionnaires realting to Ethnic and Migrant Minorities in the European Question Bank (Version 1.0)*. Ph.D. thesis, Sciences Po, CEE.

NE Sithole. 2015. *The functional viability of Indigenous African Languages in South Africa: challenges and prospects of their survival*. Ph.D. thesis, University of Zululand.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Bandeh Ali Talpur and Declan O'Sullivan. 2020. Multi-class imbalance in text classification: A feature engineering approach to detect cyberbullying in Twitter. *Informatics*, 7(4):52–74.

Robert Lawrence Trask. 2003. *Language: the basics*. Routledge.

Mihoko Wheeler. 2018. Phonetic analysis of clicks, plosives and implosives of isixhosa: A preliminary report. *Florida Linguistics Papers*, 5(2).

Marcos Zampieri. 2016. *Pluricentric languages: automatic identification and linguistic variation*. Ph.D. thesis, Universität des Saarlandes, Sao Paulo.

7. Language Resource References

Martin Puttkammer and Justin Hocking and Roald Eiselen. 2016. *NCHLT South African Language Identifier*. South African Centre for Digital Language Resources. PID https://repo.sadilar.org/handle/20.500.12185/350.

The first Universal Dependency Treebank for Tswana: Tswana-Popapolelo

Tanja Gaustad[†], **Ansu Berg**[‡], **Rigardt Pretorius**[‡], **Roald Eiselen**[†]
[†]Centre for Text Technology (CTexT), [‡]Setswana
North-West University, Potchefstroom, South Africa
{FirstName.LastName}@nwu.ac.za

Abstract

This paper presents the first publicly available UD treebank for Tswana, *Tswana-Popapolelo*. The data used consists of the 20 Cairo CICLing sentences translated to Tswana. After preprocessing these sentences with detailed POS (XPOS) and converting them to universal POS (UPOS), we proceeded to annotate the data with dependency relations, documenting decisions for the language specific constructions. Linguistic issues encountered are described in detail as this is the first application of the UD framework to produce a dependency treebank for the Bantu language family in general and for Tswana specifically.

Keywords: Dependency treebank, annotation, Bantu languages, Tswana

1. Introduction

Along with a recent push to broaden the linguistic diversity in Natural Language Processing (NLP) research (Joshi et al., 2020), there has been an increased interest in syntactic annotations for under-resourced languages from Sub-Saharan Africa resulting in treebanks for Bambara (Aplonova and Tyers, 2017; Dione, 2021), Beja (Kahane et al., 2021), Wolof, and Yoruba (Dione, 2021). The only such resource currently available for Tswana is a treebank based on Lexical Functional Grammar (Berg, 2018) consisting of phrases and simple sentences (LR Berg, 2018).

This paper describes the first publicly available Tswana[1] treebank *Tswana-Popapolelo* annotated in the Universal Dependency (UD) framework (de Marneffe et al., 2021). As a proof of concept, we chose to annotate a small data set in UD as well as document linguistic annotation issues and decisions when applying UD to Tswana so that going forward more data can be annotated more easily.

In this paper, we will focus on the building of a UD treebank for Tswana (see section 2 for the necessary background information), describing the data (section 3) and preprocessing (section 4), the annotation process (section 5) and, most importantly, issues we encountered (section 6) when trying to apply the UD framework to a novel language (family).

[1]We will be using Tswana as this is the preferred term in an international setting rather than Setswana which is generally used in South Africa (as outlined in the South African Constitution of 1996 and the Use of Official Languages Act 12 of 2012). The same decision to not use prefixes applies to the names of the other official South African languages in this article.

2. Background

2.1. Linguistic Background

Tswana (ISO-639-3 tsn) is a Bantu language spoken in the north western parts of South Africa, the eastern parts of Namibia which border on Botswana and in Botswana, where it is the national language and most of the people are first language speakers. It is one of the 12 official languages of South Africa and is spoken by 8,3% of the population (Statistics South Africa, 2023), making it the 6th most frequent home language. Next to sign language and two Germanic languages (Afrikaans and English), the official languages comprise nine Bantu languages: Four Nguni languages (Ndebele, Xhosa, Zulu, and Swati), three Sotho languages (Northern Sotho, Southern Sotho, Tswana), as well as Venḓa, and Tsonga. Tswana is classified in the South-Eastern Zone of Bantu languages. These Bantu languages are divided in language groups and Tswana is included in the Sotho language group (group S31) (Maho, 2003).

Bantu languages have a number of linguistic characteristics that make them substantially different from most Indo-European languages (van der Velde et al., 2022): all of them are tone languages; they use an elaborate system of noun classes (Katamba, 2003) and their nominal and verbal morphology is highly agglutinative and very productive (Katamba, 1993). Especially the last two characteristics are important in the context of syntactic annotation.

The selection of the orthographic writing style adopted for Tswana was influenced by historical and phonological reasons (Taljard and Bosch, 2006, 433). Phonologically, the strong homographic character of the verbal prefixes of the Sotho languages

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 55–65

(including Tswana) has led to the adoption of a disjunctive orthography regarding verbal prefixes. Nguni languages, on the other hand, have adopted a conjunctive writing system (Louwrens and Poulos, 2006). In this context, a distinction is made between *linguistic* words and *orthographic* words. For languages like English or Afrikaans, a linguistic word and an orthographic word largely coincide. For the conjunctively written languages, one orthographic word corresponds to one or more linguistic words. For disjunctively written languages like Tswana, however, several orthographic words can correspond to one linguistic word. The following example illustrates the disjunctive (Tswana) versus conjunctive (Zulu) writing styles:

Tswana	*ke a mo rata*			
	ke	a	mo	rata
	I	[pres]	him/her	love
	'I love him/her'			
Zulu	*ngiyamthanda*			
	ngi-	-ya-	-m-	-thanda
	I	[pres]	him/her	love
	'I love him/her'			

The implications of the disjunctively written verbal prefixes in the syntactic description of Tswana will be discussed in more detail in section 6.

2.2. Universal Dependencies (UD)

Universal dependencies (UD) is an international, collaborative project with two main aims: to develop a common framework describing the grammatical structure of the world's languages (de Marneffe et al., 2021) and to create treebanks for various languages applying this framework (Nivre et al., 2020). The project strives to produce cross-linguistically consistent treebanks (with language-specific extensions where necessary) describing syntactic structures as well as morphological features. This framework allows for comparisons between languages (including languages with free word order), research from a language typology perspective as well as the development of multilingual parsers. As stated in the introduction to the UD project: "The annotation scheme is based on an evolution of (universal) Stanford dependencies (de Marneffe et al., 2021), Google universal part-of-speech tags (Petrov et al., 2012), and the Interset interlingua for morphosyntactic tag sets (Zeman, 2008)."[2] The syntactic relations in UD are represented as dependency trees rather than phrase structure trees which makes the annotated data easier to use and interpret in downstream tasks. Currently, there are over 250 treebanks in more than 140 different languages available[3].

Our choice of using UD stems from the intended use of UD treebanks that benefits both the computational and linguistic research communities. As de Marneffe et al. (2021) point out, UD needs to comply with a number of (competing) criteria which include a) linguistic requirements, such as achieving a satisfactory level of annotation for linguistic analysis of individual languages and being good for highlighting structural similarities across related languages, b) computational needs, i.e. being suitable for parsing with high accuracy and supporting downstream natural language processing tasks, and, last but not least, c) pragmatic requisites, namely being suitable for rapid, consistent annotation by a human annotator and easily comprehensible by non-linguist users.

An added benefit of joining an endeavour like UD is the available infrastructure in terms of how-tos on contributing, validation scripts, support from the UD community and the visibility, availability and re-usability of the annotated data through official releases.

3. Data: Cairo CICLing Corpus

For a UD treebank to be included in an official release, it has to contain at least 20 sentences and 100 words. This can most easily be achieved by translating the 20 example sentences in the Cairo CICLing Corpus[4] to the desired language, in our case Tswana. Using these sentences has the added advantage that they contain different linguistic constructions, making it a good first test case for discussing how to annotate these constructions in the targeted language.

After procuring the 20 Cairo CICLing sentences, three Tswana native speakers[5] translated all the data without consulting each other. In a second step, the Tswana team decided on a consensus translation where consensus on the final translated sentences was attained after considering free and word-for-word translations. This was then the input to the preprocessing described next.

4. Preprocessing: Tokenisation, XPOS and UPOS

The tokenisation for the 20 translated sentences corresponds to orthographic words as used in the official orthography of Tswana (Cole, 1955; Krüger,

[2] https://universaldependencies.org/introduction.html, retrieved 12-02-2024.

[3] See https://universaldependencies.org/ for an overview.

[4] https://github.com/UniversalDependencies/cairo/blob/master/translations.txt

[5] These were graduate students paid for their time.

PRON	43	20%	ADV	12	6%
VERB	34	16%	AUX	12	6%
PART	32	15%	CCONJ	8	4%
NOUN	26	12%	SCONJ	5	2%
PUNCT	23	11%	ADJ	4	2%
PROPN	15	7%			
Total tokens: 214			Type-Token ratio: 0,47		

Table 1: Overview of UPOS tags assigned in the 20 Tswana sentences.

nsubj	25	12%	nmod	10	5%
punct	23	11%	aux	8	4%
root	20	9%	obj	8	4%
case	16	7%	compound	6	3%
expl	16	7%	fixed	6	3%
mark	11	5%	ccomp	5	2%
advmod	10	5%	xcomp	5	2%
cc	10	5%	obl	4	2%
conj	10	5%	others	21	10%

Table 2: Overview of dependency relations assigned in the 20 Tswana sentences.

2006). We will discuss ramifications and potential different choices in more detail in section 6.

An important premise when annotating universal dependencies is the presence of parts-of-speech (POS), more specifically universal POS (UPOS) (Petrov et al., 2012). The UPOS tag set contains 17 tags: 6 for open classes (nouns, verbs, etc.), 8 for closed classes (e.g. pronouns, conjunctions) and 3 for miscellaneous items (such as punctuation and symbols).

In the very limited work done on Tswana, there is not yet consensus on how to accommodate traditional Tswana POS in UPOS[6] and the application of UPOS tags is not always straightforward (Dione et al., 2023). However, there are Tswana POS taggers with more extensive tag sets (Eiselen and Puttkammer, 2014; Puttkammer et al., 2018; Malema et al., 2020; Dibitso et al., 2022). For the purposes of the Tswana UD annotations, the NCHLT tokeniser and POS tagger were used to annotate the data with detailed POS (typically referred to as XPOS in UD). This tagger includes 26 main tags, and 188 tags when including class information. The detailed POS tags were subsequently converted to UPOS tags based on a conversion table. Table 1 provides an overview of the distribution of the assigned UPOS tags in the data.

Even with the seemingly simple task of reducing the XPOS tag set to UPOS, there were a few difficult decisions during the conversion, especially to not overload one particular UPOS tag (mostly PART particle) with too many distinct XPOS categories.

The main problem concerned verbal prefixes in Tswana. As mentioned earlier, due to the disjunctive writing style, several classes of verbal morphemes are written as separate words. These morphemes are also separately tagged in the XPOS schema and include concordial morphemes (subject and object morphemes), possessive, negative, aspectual, and tense morphemes. Subject and object morphemes were treated as subject and object concords respectively in XPOS. With no direct equivalents available in UPOS, they were tagged as PRON (pronoun), while possessive concords are tagged as PART (particle). Tense and aspect morphemes (assigned the tag MORPH in XPOS) were also converted to PART (particle) in UPOS, while the negative markers were converted to ADV.

Additionally, in our detailed XPOS tag set, ideophones received their own tag IDEO as they are considered a separate word class expressing an action, manner or property through sound imitation, but not always exhibiting the same syntactic function in a sentence. However, there is no such tag in UPOS and no equivalence in other language families in the UD catalogue was found. As there were no ideophones in the Tswana Cairo CICLing sentences, the tag was not needed, but in further annotations we would consider the ADV (adverb) tag for ideophones in Tswana.

The advantage of having both XPOS and UPOS tags at our disposal during syntactic annotation is that highly ambiguous tokens (homophonic and morphosyntactically ambiguous, e.g. *ka* 'with, on, through', *go* 'copulative verb in different moods, at, to+verb, to+location') can more easily be linked correctly and that both the automatically assigned XPOS tags as well as the converted UPOS tags could be corrected if needed.

5. Annotation Process

The syntactic annotation as well as corrections to the UPOS tags was done in Arborator Grew (Guibon et al., 2020)[7]. In a first step, the annotators checked the UPOS and XPOS tags and corrected them where necessary. Then, the syntactic structure was added incrementally: start by identifying the root for each sentence, link the subject and object(s), then proceed to link and label the remaining dependencies. During the annotation, vari-

[6] Tswana has been included in the UDMorph Tagger https://lindat.mff.cuni.cz/services/teitok-live/udmorph/index.php?action=tag, but the conversion from the detailed POS tag set to UPOS has not been checked by linguists (yet).

[7] https://arboratorgrew.elizia.net/

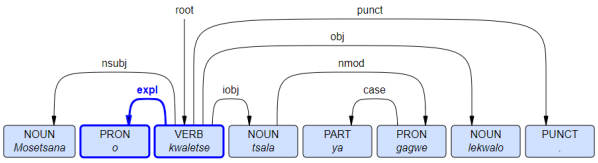


Figure 1: Sentence 1 with an overt subject.

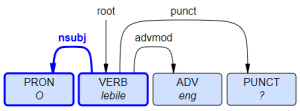


Figure 2: Sentence 2 with a covert subject.

ous options for certain syntactic constructions were discussed and documented until the annotators agreed on the preferred linking and associated dependency labels. Table 2 shows an overview of the assigned dependency relations.

The *Tswana-Popapolelo* treebank will be available in the next UD release[8], along with all language specific documentation.

6. Issues Encountered

Some of the annotations were straightforward for the current small data set, but specific linguistic idiosyncrasies of Tswana as an example of a disjunctively written Bantu language required more in-depth discussions on how to use existing UD relations. These are detailed below.

6.1. Disjunctive Orthography and Verbal Constructions in Tswana

As has been described in section 2.1, a disjunctive writing style has been adopted for Tswana. Especially for verbal prefixes, this means that a large number of orthographic tokens preceding the verb would in traditional linguistics be seen more as morphemes rather than "proper" words. The proper identification of words is generally taken as an imperative preprocessing step for syntactic description. In this regard, the disjunctively written verbal prefixes of Tswana cause orthographic tokens which are part of linguistic words (Taljard and Bosch, 2006) and therefore compromise the lexical integrity of verbs. Tswana verbal prefixes carry inflectional information while suffixes carry inflectional as well as derivational information (Krüger, 2006, 268). These verbal prefixes also carry both morphological and syntactic information which makes it difficult to assign UPOS tags and syntactic relations to them.

One obvious solution to address this problem is to adjust the tokenisation to reflect linguistic words, rather than orthographic words. Although this would certainly simplify and more closely align the data with the UD framework, this would also reduce the granularity and informativeness of the treebank. With this in mind, we opted to annotate the relations between all orthographic words. More details on the implication of tokenisation is provided in section 6.5.

In the UD annotation of Tswana, the disjunctively written verbal prefixes are linked to the verbal root via arcs. We will now describe how the different parts of verbal constructions have been handled.

6.1.1. Subject Concords

In instances where an overt subject is realised in a sentence, the subject concord is an agreement marker which marks the relation between the overt subject and the verb. In these cases we opted for the `expl` relation. This relation is used in UD for phenomena such as clitic doubling (e.g. in Romance languages) or the doubling of a lexical nominal and a pronominal clitic (e.g. in Greek and Bulgarian). Even though subject concords are not the same as clitics, they behave in a similar fashion in that they are a type of "pronominal" copy without its own semantic role. An example for Tswana can be seen in figure 1 of sentence 1.[9]

(1) Mosetsana **o** kwaletse
girl she[SubjConc] write[appl-perf]
tsala ya gagwe lekwalo
friend of her letter
'The girl wrote a letter to her friend.'

In instances where the overt subject is not realised, the (covert) subject concord acquires a

[8] https://github.com/UniversalDependencies/UD_Tswana-Popapolelo

[9] All figures were produced with http://www.let.rug.nl/kleiweg/conllu/.

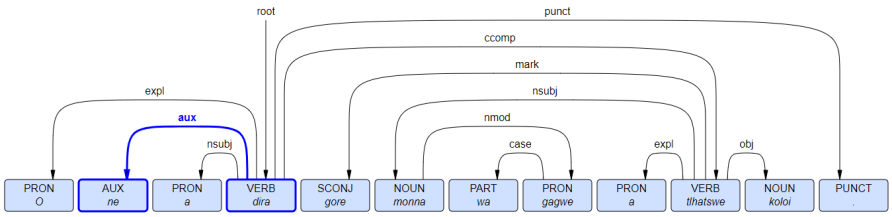


Figure 3: Sentence 3 showcasing the use of the `aux` relation.

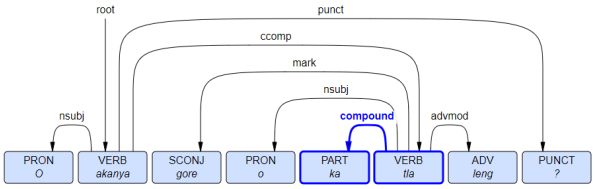


Figure 4: Sentence 4 showcasing the use of the `compound` relation for certain TAME morphemes.

pronominal status (as it would be a pronoun in translation) and becomes the actual subject, hence is annotated with the `nsubj` relation as shown in figure 2 of sentence 2.

(2) **O** lebile eng
you[SubjConc] see[perf] what
'What are you looking at?'

6.1.2. Auxiliary Verbs

An auxiliary verb enriches the meaning of the complementary main verb, copulative verb or another auxiliary verb phrase and can add semantic information regarding the mood, tense, aspect and/or polarity of a verb. It also adds information on the progression or completion of an action: It expresses a certain type of duration of the action or it expresses the logical time at which the action is executed. For example, the auxiliary verb *ne* expresses a relative past tense indicating that the action was taking place or had taken place at some point in the past. If the complementary verb is in the present tense then it indicates an action that is incomplete and continuing at a certain moment in the past. If the complementary verb is in the perfect it indicates that the action had been completed at the point of reference (Pretorius, 1997; Krüger, 2013a).

In UD, auxiliary verbs are a closed class that cannot have any children. The `aux` relation is used in Tswana to indicate the relation between a verb and the preceding auxiliary verb, as with other languages. However, we encountered the issue of auxiliaries taking a (doubled) subject concord as a dependent. In sentence 3, the subject concord occurs twice: once (realised as *o*) with the auxiliary *ne* and once (realised as *a*) with the verb *dira*[10], but both referring to the subject 'she' and both needed for the sentence to be grammatical.

(3) **O** **ne**
she[SubjConc] aux[past-indef]
a dira gore monna wa
she[SubjConc] make that husband of
gagwe a tlhatswe koloi
her he[SubjConc] wash[pass] car
'She made her husband wash the car.'

At this stage, we have chosen to annotate the subject concord with the auxiliary verb with a `expl` relation, while the subject concord with the main verb becomes the `nsubj` and the relation between the auxiliary and the main verb is tagged `aux` as can be seen in figure 3.

6.1.3. TAME Morphemes

The disjunctively written Tense-Aspect-Mood-Evidentiality (TAME) morphemes in the morphological structure of a verb always occur in a fixed order and are not morphosyntactically flexible. For the TAME morphemes including the present tense morpheme *a*, the progressive morpheme *sa* 'still', the potential morpheme *ka* 'can, may' and the future tense morpheme *tla* 'will, shall' the `compound` relation is applied to express that this is a combination of lexemes that morphosyntactically behave as single words. See the example in sentence 4 and figure 4.

(4) O akanya gore
you[SubjConc] think that
o ka **tla** leng
you[SubjConc] can come when
'When do you think you can come?'

[10]The meaning of the auxiliary verb *ne* requires the consecutive form of the subject agreement morpheme following it.

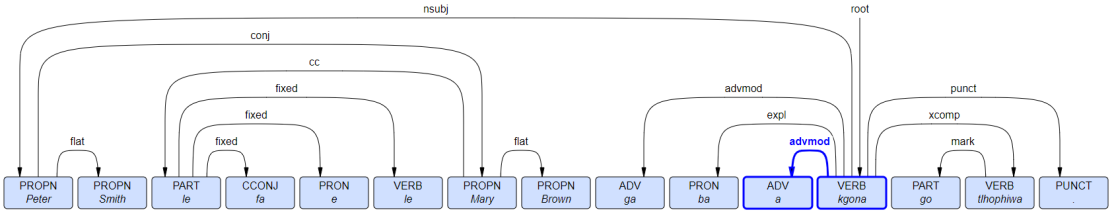


Figure 5: Sentence 5 showcasing the use of the `advmod` relation for negation TAME morphemes.

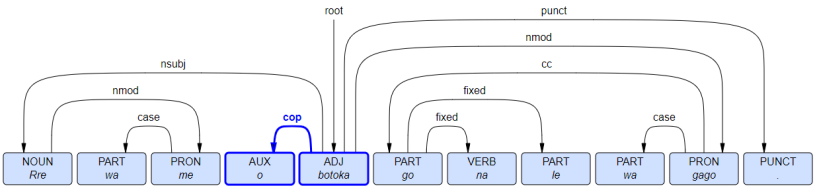


Figure 6: Sentence 6 with a describing copulative verb.

For the negative morphemes *ga*, *sa*, *se*, we have opted to use the combination of ADV and `advmod`, parallel to English, as illustrated in sentence 5 and figure 5.

(5) Peter Smith le fa e le Mary Brown **ga**
Peter Smith neither nor Mary Brown not
ba **a** kgona go
they[SubjConc] not can to[InfMarker]
tlhophiwa
select[pass]
'Neither Peter Smith nor Mary Brown could be selected.'

6.2. Copulatives

A copula is the relation of a function word used to link a subject to a nonverbal predicate. In Tswana three types of copulative verbs are distinguished:

- identifying copulative: identifies a subject with regards to type, status or profession or to predicate the existence or presence of a thing, e.g. *Lekwalo lê ke la gago* 'This letter is yours';
- describing copulative: establishes some quality, characteristic or state of a subject, or its situation or locality, e.g. *Ditlhako tsa me di dintšha* 'My shoes are new';
- associative copulative: expresses the idea of the English have or be with and indicates possession or association, e.g. *Sediba sê se na le metsi* 'This well has water'.

The morphological structure of these verbs may include tense, aspect, mood and polarity information.

When the verb in Tswana is an identifying or describing copulative verb, the root of the clause is the complement of the copulative verb. These two types of copulative verbs are POS tagged as AUX, and the `cop` relation is used between the root and the preceding copulative verb. See sentence 6 and figure 6 for an example of a describing copulative.

(6) Rre wa me **o** **botoka** go na le wa
father of me is[cop] cooler than of
gago
you
'My dad is cooler than yours.'

In the case of an associative copulative verb, the root of the clause is the copulative verb. The associative copulative verbs in Tswana are POS tagged as VERB, and the `obj` relation is used between the root and the complement that follows it, as showcased in sentence 7 and figure 7. This analysis differs from traditional Tswana linguistic descriptions (Cole, 1955; Krüger, 2006, 2013b).

(7) Ga ba **na** **kakanyo** epe
not they[SubjConc] have idea none
gore e kwadilwe ke
that it[SubjConc] write[perf-pass] by
mang.
who
'They have no idea who wrote it.'

6.3. Use of the `mark` Relation

Conjunctions that mark a clause as subordinate to another clause are annotated as `mark` in UD. In Tswana, the marker is an introductory member of a clause that includes an action in the subjunctive or participial mood. For the subjunctive, the conjunction *gore* 'that, so that' is used, as shown in sentence 8 and figure 8. For the participial, a conjunction such as *fa* 'as, while, when, if', *le fa* 'even if, although, while' and *ka* 'since' are used.

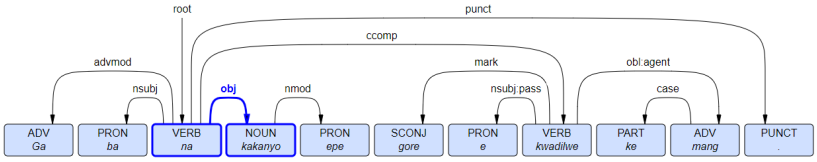


Figure 7: Sentence 7 with an associative copulative verb.

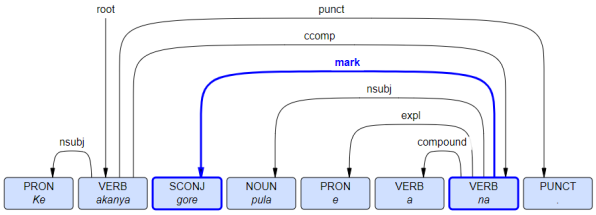


Figure 8: Sentence 8 showcasing the use of the `mark` relation for a subordinate clause.

(8) Ke akanya **gore** pula
I[SubjConc] think that rain
e a na
it[SubjConc] [pres] falls
'I think that it is raining.'

The `mark` relation is also used in Tswana for infinitive verbs, analogue to English and German, for example *go tlogela* 'to stop' where the marker is *go* 'to'. See sentence 9 and figure 9 for an illustration.

(9) O ne
he[SubjConc] aux[past-indef]
a leka **go** **tlogela**
he[SubjConc] try to[InfMarker] quit
go goga le go
to[InfMarker] smoke and to[InfMarker]
nwa
drink
'He tried to stop smoking and drinking.'

Furthermore, `mark` is used in a relative clause where the qualificative particle is the marker as seen in sentence 10 and figure 10: The qualificative particle always agrees with a specific noun class in Tswana, for example in *e kgolo* '[part] big' the marker is the qualificative particle *e* that indicates noun class 9 agreement.

(10) A Iguazu ke naga **e**
[InterPart] Iguazu is[cop] country [part]
kgolo kgotsa ke **e** nnye
big or is[cop] [part] small
'Is Iguazu a big or a small country?'

6.4. Interrogative Particle *a*

In Tswana there is an interrogative particle *a* added at the beginning of a sentence to change an indicative sentence to an interrogative one. After consultation with the UD community, we have decided to assign the UPOS tag PART (particle) to *a* as well as link it directly to the root of the sentence with a `discourse` relation (following the Latin example of *ne*). As this particle works on a more pragmatic level, the `discourse` relation "used for interjections and other discourse particles and elements (which are not clearly linked to the structure of the sentence, except in an expressive way)" as described in the UD overview of relations[11] seemed the best choice. Sentence 11 and figure 11 show an example of this for Tswana.

(11) **A** o batla
[InterPart] you[SubjConc] want
go tsamaya?
to[InfMarker] leave/go
Do you want to leave/go?

6.5. Tokenisation in Tswana

An issue that we definitely have not solved yet and that is connected to the previous section 6.1 is the tokenisation of Tswana. Traditionally, computational analyses for disjunctively written South African Bantu languages, i.e. Northern Sotho, Southern Sotho, Tswana, Venda and Tsonga, have been done on orthographic words as then no conversions are needed from the original text. The implications of choosing to use orthographic words rather than linguistic words, however, will be felt at various levels when working on the syntactic analysis of Tswana applying UD dependencies:

- When doing annotation: Working on the orthographic word means more time and effort will be spent on getting the UPOS as well as

[11] https://universaldependencies.org/u/dep/all.html#al-u-dep/discourse.

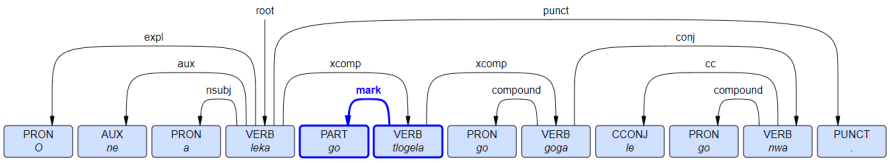


Figure 9: Sentence 9 showcasing the use of the `mark` relation in infinitives.

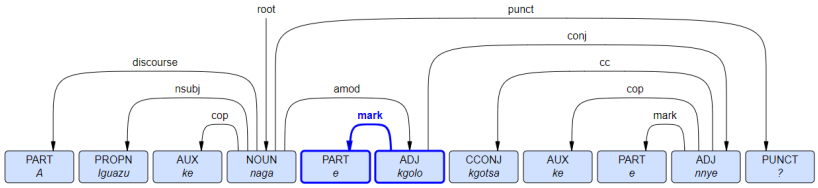


Figure 10: Sentence 10 showcasing the use of the `mark` relation for qualificative particles.

the dependency relations right (both pertaining more to the syntactic level). With linguistic words, the syntactic structure becomes more straight forward (simpler?), but at the same time more care needs to be given to adding morphological information to retain the necessary detail to be able to disambiguate.

- From a computational linguistics view point: Specifically in the UD framework, if Tswana text is analysed using orthographic words, the resulting annotations make it more directly comparable with European languages and the syntactic annotations will be more diverse and informative. On the other hand, using linguistic words will put more emphasis on the similarities with other, especially conjunctively written Bantu languages, but the syntactic structure will be simpler as a lot of information will only be contained on the morphological (sub-word) level.

- In relation to linguistic analyses: In traditional (structural) grammatical descriptions, the left hand boundary of Tswana verbs is considered to be the first prefix of such a verb, even if it is written disjunctively. This implies that verbs such as *ke a mo rata* in 2.1 would be tokenised as one word, namely a verb. This verb would constitute a sentence in itself and would be the predicate of the sentence. The syntactic analysis of the sentence would thus not indicate the pronominal value of the subject and object concords so as to indicate that the sentence contains a subject and object (Taljard and Bosch, 2006; Louwrens and Poulos, 2006; Krüger, 2006; Cole, 1955; Pretorius et al., 2015). In later descriptions (Berg, 2018), the lexical integrity of the verb is maintained but the argument status of these concords is indicated on the functional level.

So, if we were to decide to "attach" verbal prefixes to the verb, the original structure in 12 based on orthographic words would change to the representation in 13.

(12) **Ga ke a kgona**
not I[SubjConc] [pres] able
go tshwarelela ka gore
to[infMarker] keep up because
o ne
he[SubjConc] aux[past-indef]
a taboga ka lebelo thata
he[SubjConc] run with speed much

(13) **[Ga ke a kgona]** go tshwarelela ka gore
[I wasn't able] to keep up because
[o ne] **[a taboga]** ka lebelo thata
[he aux] [he ran] with speed much
'I wasn't able to keep up, because he ran too fast.'

We feel more work is needed to explore where to draw the boundaries when "attaching" verbal prefixes as well as to fully understand the consequences of such an approach.

7. Conclusion and Future Work

This paper contains the description of the first publicly available UD treebank for Tswana, based on the 20 translated Cairo CICLing sentences. The resulting treebank shows that this was a successful first endeavour to apply UD to a Bantu language and forms the basis for further annotation of Tswana to a more extensive set. The main benefit of starting with such a small data set is that many of the most problematic annotations can be discussed in detail, and the corresponding outcomes can be documented without needing a substantial reannotation of the data at a later stage. As would be expected, not all issues have been resolved yet and

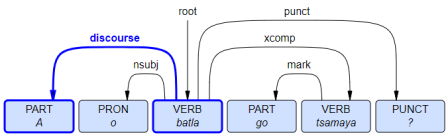


Figure 11: Sentence 11 with an interrogative particle.

some decisions had to be made on how to best apply the existing framework to a novel language with unique linguistic characteristics. We hope the detailed report on the issues encountered will also help others when annotating new Sotho and Bantu languages in UD.

With the *Tswana-Popapolelo* treebank now available, we plan to annotate extra data with the help of student assistants. The current annotations are based on our understanding of the literature and feedback we received from the UD community, but the choices made thus far will definitely be further refined and the available annotated data for Tswana will be expanded by adding it to *Tswana-Popapolelo*. This includes experimenting with different tokenisation strategies for the same data to study the repercussions on the dependency analyses.

Once a larger set of treebank data is available, we will also train automatic parsers to pre-annotate data to assist and simplify the annotation process. Ultimately we aim to have enough data to train accurate full dependecy parsers, including XPOS, UPOS, lemma and morphological taggers, while at the same time leveraging the work of others that use UD treebanks to train various NLP tools.

8. Acknowledgements

We would like to thank Kevin Mavalela and Kaboentle Maibi for their contribution to the translations and initial discussions and annotations. Furthermore, we are grateful to the UD community for their responses to our queries.

9. Bibliographical References

Ekaterina Aplonova and Francis M. Tyers. 2017. Towards a dependency-annotated treebank for Bambara. In *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, pages 138–145, Prague, Czech Republic.

Ansu Berg. 2018. *A computational syntactic analysis of Setswana*. Ph.D. thesis, North-West University, Potchefstroom, South Africa.

Desmond T. Cole. 1955. *An introduction to Tswana grammar*. Longman, Cape Town.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Mary Dibitso, Pius A. Owolawi, and Sunday O. Ojo. 2022. An hybrid part of speech tagger for Setswana language using a voting method. In *International Conference on Intelligent and Innovative Computing Applications*, pages 245–253.

Cheikh M. Bamba Dione. 2021. Multilingual dependency parsing for low-resource African languages: Case studies on Bambara, Wolof, and Yoruba. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 84–92, Online. Association for Computational Linguistics.

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba O. Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. 2023. MasakhaPOS: Part-of-speech tagging for typologically diverse African languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers*, pages 10883–10900.

Roald Eiselen and Martin J. Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 3698–3703.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. 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*, pages 6282–6293, Online. Association for Computational Linguistics.

Sylvain Kahane, Martine Vanhove, Rayan Ziane, and Bruno Guillaume. 2021. A morph-based and a word-based treebank for Beja. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 48–60, Sofia, Bulgaria. Association for Computational Linguistics.

Francis Katamba. 1993. *Morphology*. Springer, New York.

Francis Katamba. 2003. Bantu nominal morphology. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, pages 103–120. Routledge, London & New York.

Caspar J.H. Krüger. 2006. *Introduction to the morphology of Setswana*. Lincom Europe, München.

Caspar J.H. Krüger. 2013a. *Setswana syntax: a survey of word group structures: Volume 1*. Lincom Europe, München.

Caspar J.H. Krüger. 2013b. *Setswana syntax: a survey of word group structures: Volume 2*. Lincom Europe, München.

Louis J. Louwrens and George Poulos. 2006. The status of the word in selected conventional writing systems - the case of disjunctive writing. *Southern African Linguistics and Applied Language Studies*, 24(3):389–401.

Jouni Maho. 2003. A classification of the Bantu languages: an update of Guthrie's referential system. In Derek Nurse and Gérard Philippson, editors, *The Bantu Languages*, pages 639–651. Routledge, London & New York.

Gabofetswe Malema, Boago Okgetheng, Bopaki Tebalo, Moffat Motlhanka, and Goaletsa Rammidi. 2020. Complex Setswana parts of speech tagging. In *Proceedings of the first workshop on Resources for African Indigenous Languages (RAIL)*, pages 21–24.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA).

Laurette Pretorius, Biffie Viljoen, Ansu Berg, and Rigardt Pretorius. 2015. Tswana finite state tokenisation. *Language Resources and Evaluation*, 49(4):831–856.

Rigardt Pretorius. 1997. *Auxiliary Verbs as a Subcategory of the Verb in Tswana*. Ph.D. thesis, PU for CHE, Potchefstroom, South Africa.

Martin Puttkammer, Roald Eiselen, Justin Hocking, and Frederik Koen. 2018. NLP web services for resource-scarce languages. In *Proceedings of ACL 2018, System Demonstrations*, pages 43–49, Melbourne, Australia. Association for Computational Linguistics.

Department of Statistics South Africa. 2023. Census 2022: Statistical release. Technical report, Department of Statistics, Republic of South Africa.

Elsabé Taljard and Sonja E. Bosch. 2006. A comparison of approaches to word classtagging: Disjunctively vs. conjunctively written Bantu languages. *Nordic Journal of African Studies*, 15(4):428–442.

Mark van der Velde, Koen Bostoen, Derek Nurse, and Gérard Philippson, editors. 2022. *The Bantu Languages*, 2nd edition. Routledge, London & New York.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

10. Language Resource References

LR Berg, Ansu. 2018. *Setswana Test suite and Treebank*. South African Centre for Digital Language Resources (SADiLaR). PID https://hdl.handle.net/20.500.12185/478.

Adapting Nine Traditional Text Readability Measures into Sesotho

Johannes Sibeko, Menno van Zaanen
Nelson Mandela University, South African Centre for Digital Language Resources
University way, Summerstrand, Port Elizabeth, Internal Box 340, Private bag X6001, Potchefstroom
johanness@mandela.ac.za, menno.vanzaanen@nwu.ac.za

Abstract

This article discusses the adaptation of traditional English readability measures into Sesotho, a Southern African indigenous low-resource language. We employ the use of a translated readability corpus to extract textual features from the Sesotho texts and readability levels from the English translations. We look at the correlation between the different features to ensure that non-competing features are used in the readability metrics. Next, through linear regression analyses, we examine the impact of the text features from the Sesotho texts on the overall readability levels (which are gauged from the English translations). Starting from the structure of the traditional English readability measures, linear regression models identify coefficients and intercepts for the different variables considered in the readability formulas for Sesotho. In the end, we propose ten readability formulas for Sesotho (one more than the initial nine; we provide two formulas based on the structure of the Gunning Fog index). We also introduce intercepts for the Gunning Fog index, the Läsbarhets index and the Readability index (which do not have intercepts in the English variants) in the Sesotho formulas.

Keywords: Text Readability, Sesotho, Low-resource language

1. Introduction

The reports from the Progress in International Reading Literacy Study (PIRLS) show consistent subpar performance among learners reading in South African indigenous languages (Roux et al., 2021). In the PIRLS standards, learners who perform below the 400-point benchmark, struggle to extract fundamental information from the text, making it challenging for them to respond to even the simplest questions. Regrettably, at least 81% of learners in the South African indigenous languages have been performing below the 400-point benchmark (Roux et al., 2021). As a result, such performance hinders the achievement of inclusive and equitable quality education in essentially all high school subjects as learners cannot access information from written sources. Steps need to be taken to address this literacy challenge as highlighted by the fourth of United Nations' (UN) seventeen Sustainable Development Goals, which focuses on the importance of ensuring inclusive and equitable quality education and promoting lifelong learning opportunities for all.

A possible solution to low literacy levels is to make sure children learn to read properly, which can only be attained through practising reading (van Bergen et al., 2018). In other words, learners need to read in order for their reading skills to improve. One way of igniting the desire to read is providing learners with both opportunities to select texts and reading time (Rasheed, 2023). According to Rasheed (2023), learners who have the autonomy to choose their reading materials tend to perform better than those who are assigned texts. However, it is essential to note that a poor choice of reading materials can hinder the development of reading skills when the texts are not well-matched to the reader's level of proficiency (Mohammed et al., 2023). Keeping this in mind, it becomes evident that education stakeholders require a tool to assess the readability of texts to enable the identification of texts that align with the reader's reading ability level. The development of readability measures for the different indigenous languages of South Africa will allow for objective measurements of text readability.

Note that the indigenous languages of South Africa are low-resourced. As such, the choice of approaches to the exploration of text readability is somewhat limited. Here, we propose the use of traditional readability measures that focus on shallow text properties (Van Oosten et al., 2010; Zamanian and Heydari, 2012).

Despite a longstanding research interest in readability assessment, traditional readability measures have not been tailored for South African indigenous languages (Leopeng, 2019). The lack of text readability measures for South African indigenous languages so far has led to the use of (unmodified) English readability measures for readability analyses in indigenous South African languages such as isiZulu (Land, 2015), isiXhosa (Carel, 2019; Leopeng, 2019), and Sesotho (Krige and Reid, 2017; Reid et al., 2019). Recently, Sibeko (2023) reports attempts to develop text readability measures for Sesotho. Their article focuses on the basic language resources for Sesotho required to develop the readability measures. However, they

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 66–76

do not tackle the actual development of readability measures.

In this article, we focus on the development of readability measures for Sesotho and not all twelve official languages of South Africa. Even though a similar approach may be applied to the other languages as well, sign language, one of the twelve official languages, may require a different approach. Overall, we address the research question:

> How can traditional readability measures be effectively modified and adapted to suit the specific characteristics of Sesotho?

To answer this question, we adapt traditional readability measures to Sesotho using English as a high-resource helper language for the low-resource Sesotho. The underlying assumption is that texts that are easy to read in Sesotho will also be easy to read when translated into English and difficult Sesotho text will be translated into difficult English texts. First, the background of this investigation is presented in Section 2, then the methodology is described in Section 3, followed by the evaluation in Section 4. Finally, we present our discussion and conclusions in Section 5.

2. Background

2.1. An overview of Sesotho

Sesotho is a language spoken in Southern Africa. It is one of the two official languages in Lesotho (Government of Lesotho, 1993), one of the twelve official languages in South Africa (Republic of South Africa, 2023), and one of the marginalised official languages in Zimbabwe (Parliament of Zimbabwe, 2021). Furthermore, Sesotho is spoken in Zambia, Namibia, and Botswana. At least more than ten million people use Sesotho on a daily basis. It is used and taught in both basic and higher education sectors.

Sesotho has at least six recognised dialects, namely, the Sekwena, Sekgolokwe, Serotse, Setlokwa, Sephuthi, and Setaung (Kula and Marten, 2008; Mohasi and Mashao, 2005; Nhlapo, 2021). Of these dialects, Sekwena was promoted and has thus become the standard of writing in Sesotho (Nakin, 2009; Sekere, 2004). Moreover, there are at least two officially recognised orthographies for Sesotho, namely, the South African and the Lesothan orthographies (Makutoane, 2022; Setaka, 2018; Setaka and Prinsloo, 2020; Sibeko, 2022). The research described in this article is based on texts that are written using the South African Sesotho orthography.

2.2. Traditional Readability Measures

In this article, we explore nine traditional English readability measures for adaptation to Sesotho. These measures include the Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Flesch-Reading Ease (FRE) (Flesch, 1948, 1974), Simple Measure of Gobbledygook (SMOG) (Mc Laughlin, 1969; Zhou et al., 2017), and Gunning Fog Index (GFI) (Gunning, 1952, 1969), which rely on syllable-related information, as well as the Coleman-Liau index (CLI) (Coleman and Liau, 1975), Automatic Readability index (ARI) (Kaur et al., 2018; Smith and Senter, 1967), Readability index (RIX), and Läsbarhets index (LIX) (Björnsson, 1968; Björnsson, 1983) measures which are based on word-length information. Finally, we also explore the Dale-Chall index (Dale and Chall, 1948) which draws from a list of commonly used words. The formulas of each of these measures, as well as the type of output, are presented in Table 1.

The general approach of the syllable-based measures is to consider the number of syllables in each word and process the results in measure-specific ways. The FKGL and the FRE process syllable information by evaluating the number of syllables per word while the SMOG and the GFI measures exclude "simple" words with two or fewer syllables, thereby focusing only on words with three or more syllables.

Given that the number of syllables in long words is language-dependent, we suspected that the English requirement of 3+ syllables may not be indicative of long words as measured by the number of syllables per word in Sesotho. For instance, in a similar study, Kusec et al. (2002) adjusted the minimum syllables counted from the English helper language to the low-resource language, Croatian. They compared the top 100 frequently used words in English and Croatian to determine the differences between syllable counts in the two languages in order to determine the number of syllables that are typical in Croatian long words. In the end, they adjusted the requirement for polysyllabic words to 4+ syllables. We consider both 3+ and 4+ syllable long words in our experiments.

In addition to syllable information, word length and sentence lengths are also common features used in the measures as is evident in Table 1. Orthographic word length, that is, the lengths of words as measured by the number of letters per word (Ziegler et al., 2001), has been a topic of interest in language studies, with research indicating variations across languages and over time. For instance, Bochkarev et al. (2015) investigate the evolution of word lengths in English and Russian as observed through e-libraries, Google Books, and Google Ngram Viewer. Their findings indicate an increase in the average length of words in both

Measure	Formula	Output
FKGL	$= 0.39(\frac{\#tokens}{\#sentences}) + 11.8(\frac{\#syllables}{\#tokens}) - 15.59$	grade
FRE	$= 206.835 - 1.015(\frac{\#tokens}{\#sentences}) + 84.6(\frac{\#syllables}{\#tokens})$	level
SMOG	$= 3.1291 + 1.043\sqrt{\#polysyllabicwords * (\frac{30}{\#sentences})}$	grade
GFI	$= 0.4[(\frac{\#tokens}{\#sentences}) + 100(\frac{\#complex-words}{\#words})]$	grade
CLI	$= 0.0588(\frac{\#letters}{\#samples}) - 0.296(\frac{\#sentences}{\#samples}) - 15.8$	grade
ARI	$= 4.7(\frac{\#letters}{\#words}) + 0.5(\frac{\#words}{\#sentences}) - 21.43$	grade
RIX	$= \frac{\#longwords}{\#sentences}$	grade
LIX	$= (\frac{\#words}{\#sentences}) + [\frac{\#longwords}{\#words} * 100]$	grade
DCI	$= 0.0496(\frac{\#words}{\#sentences}) + (\frac{\#difficultwords}{\#words} * 0.1579) + 3.6365$	grade

Table 1: Selected classical readability measures (Flesch-Kincaid Grade Level (FKGL), Flesch-Reading Ease (FRE), Simple Measure of Gobbledygook (SMOG), Gunning Fog Index (GFI), Coleman-Liau index (CLI), Automatic Readability index (ARI), Readability index (RIX), Läsbarhets index (LIX), Dale-Chall index (DCI)), corresponding formulas, and type of output.

languages, with English increasing from 4.4 letters per word in the year 1700 to 4.6 in the year 2000. Additionally, they note that these numbers were reported differently in other studies where the average length of words in English was 5.1 letters per word while that of Russian was slightly higher at 5.28 letters per word (Bochkarev et al., 2015). According to Hefer (2013) words in Sesotho are on average almost a full character shorter than in English. Conversely, Loukatou (2019) indicates an average word length of 4.24 for Sesotho and a lower average of 3.02 letters per word for English in their over-segmentation corpus.

3. Methodology

According to De Clercq et al. (2014), there are at least three steps to describe when developing readability measures. Those are (i) the development of a readability corpus, (ii) describing a methodology, and (iii) undertaking the prediction tasks (François and Fairon, 2012; Collins-Thompson, 2014). We structure the discussion of our methodology for adapting the traditional readability measures into Sesotho using these three steps below.

3.1. Step 1: A readability corpus

Within the context of indigenous languages of South Africa, including Sesotho, the unavailability of readily annotated corpora with readability levels highlights the need to develop new corpora or repurpose existing corpora to train readability measures. In this context, Sibeko and Van Zaanen (2021) suggest the use of examination texts for the creation of readability corpora for South African indigenous languages.

For our study, we employ Sibeko's (2024) readability corpus of Sesotho-English translations. This corpus includes document-level parallel translations of 80 Sesotho reading comprehension and summary writing texts sourced from the grade 12 examination corpus (Sibeko and Van Zaanen, 2023). For texts produced after 2011, the English translations are essentially back translations as the texts were originally translated from English to Sesotho for exam purposes. Note that the Sesotho exam texts indicate that the original source is in English, but they do not indicate exactly where the English texts can be found (hence the back translation process is applied).

The corpus comprises 13,793 words, consisting of 6,040 types, with an average sentence length of 17.73 words in Sesotho. Additionally, the English translations include 12,005 words with 6,130 types, featuring an average sentence length of 15.75 words. The examination texts span from the year 2009 to 2019.

3.2. Step 2: A methodology

The overall methodology consists of three steps. First, we extract relevant text features from Sesotho texts. Second, we use the English translations that correspond to the Sesotho texts to determine readability levels for the texts using traditional readability measures. With this approach, we follow El-Haj and Rayson (2016) who illustrate that the readability of texts in a higher-resourced language can be utilized as a benchmark for the estimation of the readability of texts in a low-resource language. Similarly, we

align the distribution of readability levels in Sesotho with those observed in English translations. Third, we use linear regression models to determine the impact of the text features from the Sesotho texts on the overall scores of the different readability measures computed on the English translations.

To provide some additional insight into the impact of the different text features, we examine the text characteristics employed in traditional readability measures. The following brief discussion outlines some of the text features considered in this article.

3.2.1. Word lengths by letters

There are two main concerns with average word lengths in Sesotho. On the one hand, as an agglutinative language, words may be expected to be relatively long in Sesotho (Blanchard, 2011). On the other hand, monosyllabic words which may comprise between one and four letters (and especially single-letter words) may result in shorter averages for Sesotho texts (Messerschmidt et al., 2003). Furthermore, overall text word length by the letters may be affected by the use of subject concords in Sesotho. Within our dataset, English words exhibit an average of 4.34 letters per word, while Sesotho words demonstrate an average of 4.07 letters per word. Nonetheless, given that the average word length in Sesotho is relatively similar to that of English, we follow the English guideline for the LIX and RIX measures and thus consider words with more than six letters as long words.

3.2.2. Word length by syllables

Polysyllabic words refer to words with more than one syllable. However, the traditional measures used in this research, particularly the SMOG and the GFI measures consider only words with three or more syllables as polysyllabic, foggy, and complex. Within our data set, the English words exhibit an average of 1.26 syllables per word while the Sesotho texts demonstrate 2.0 syllables per word. Sesotho words tend to have more syllables than English words. As such, although we define polysyllabic words (as used in the different metrics) as words with three or more syllables, we also investigate the possibility of increasing the minimum syllables in polysyllabic words to words with four syllables.

3.2.3. Common words

The DCI measure is based on the assumption that there are words that are commonly used and should therefore be easy to read. According to this method, words that do not appear on the list of frequently used words are considered difficult. For our experiment, we use the list of common Sesotho words compiled by Sibeko and De Clercq (2023). We need to use this list with caution, however, since it was not derived from educational texts. Unfortunately, we are not aware of any other word lists available for use in this context.

3.2.4. Samples

Some formulas, like the DCI and CLI measures, require sampling of small amounts of text. As the texts in these experiments are relatively short, we forgo the sampling steps. In this way, for instance, the number of sentences in the CLI formula refers to all sentences in the text instead of a small set of sampled sentences. As can be observed in Table 1, the CLI formula focuses only on word lengths as counted in letters, and sentences in the whole text (for both the English and Sesotho formulas).

3.3. Step 3: Prediction tasks

3.3.1. Correlations

Before we develop text readability measures for Sesotho, we first investigate the interrelationships among the different textual features that underpin the readability measures. This exploration provides insights into the nature of Sesotho text features.

The exploration of the interrelationships between the text features used in the traditional readability formulas was computed using the Pearson correlation measure. The outcomes of these correlations are presented in Table 2. Note that the Labels V1-16 are used to represent the features in columns 1 to 16. Notably, all correlations are significant with $p < .05$.

The examination of Sesotho text features through correlation analysis reveals interesting findings. For example, perfect alignments are uncovered between word and syllable counts, as well as between syllable and letter counts. This suggests a consistent and predictable relationship between these features in that more syllables will result in longer words. Furthermore, strong positive correlations emerge, highlighting the association between syllables per word and the frequency of polysyllabic words, while negative correlations indicate that an increase in letter counts per word may result in fewer sentences, words, and long words.

We also investigated the correlations between syllable-based formulas and syllable-related text features. The findings in Table 3 reveal weak negative correlations between the number of syllables and the scores of the FKGL, the GFI, and the SMOG index. This observation suggests that syllable counts in Sesotho align with those in English, indicating that texts with higher syllable counts are likely to be more challenging to read.

As mentioned earlier, we also considered modifying the criteria for defining polysyllabic words

Feature	Label	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
number of sentences	V1	1.00	.87	.91	.72	.86	.86	.82	.84	.54	-.34	-.29	.29	-.48	.58	-.34
number of words	V2	.87	1.00	.93	.92	.98	1.00	.97	.99	.10	-.24	-.24	.24	-.07	.43	-.21
number of difficult_words	V3	.91	.93	1.00	.84	.94	.93	.91	.92	.28	-.20	-.18	.17	-.24	.71	-.21
number of long_words	V4	.72	.92	.84	1.00	.91	.95	.97	.96	-.05	.11	.10	-.10	.08	.38	.17
number of types	V5	.86	.98	.94	.91	1.00	.98	.96	.98	.11	-.20	-.20	.19	-.09	.50	-.18
number of syllables	V6	.86	1.00	.93	.95	.98	1.00	.99	1.00	.08	-.15	-.15	.15	-.05	.44	-.13
number of polysyllables	V7	.82	.97	.91	.97	.96	.99	1.00	.99	.05	-.04	-.03	.03	-.03	.45	-.02
number of letters	V8	.84	.99	.92	.96	.98	1.00	.99	1.00	.06	-.12	-.13	.13	-.03	.43	-.10
ratio of sentences per word	V9	.54	.10	.28	-.05	.11	.08	.05	.06	1.00	-.24	-.16	.16	-.95	.48	-.32
ratio of letters per word	V10	-.34	-.24	.20	.11	-.20	-.15	-.04	-.12	-.24	1.00	.93	-.93	.21	-.01	.89
ratio of syllables per word	V11	-.29	-.24	.18	.10	-.20	-.15	-.03	-.13	-.16	.93	1.00	-1.00	.14	.05	.88
ratio of words per syllable	V12	.29	.24	.17	-.10	.19	.15	.03	.13	.16	-.93	-1.00	1.00	-.13	-.05	-.89
ratio of words per sentence	V13	-.48	-.07	.24	.08	-.09	-.05	-.03	-.03	-.95	.21	.14	-.13	1.00	-.44	.31
ratio of difficult_words per word	V14	.58	.43	.71	.38	.50	.44	.45	.43	.48	-.01	.05	-.05	-.44	1.00	-.08
ratio of long_words per word	V15	-.34	-.21	.21	.17	-.18	-.13	-.02	-.10	-.32	.89	.88	-.89	.31	-.08	1.00

Table 2: The correlation of text features used in the readability measures computed from the Sesotho texts.

by exploring a potential increase from three to a minimum of four syllables (4+ syllables). Our findings reveal that maintaining a minimum of three syllables consistently demonstrates stronger correlations with readability scores compared to a minimum of four syllables. Consequently, for Sesotho, we also consider only 3+ syllables as in the original English formulas.

3.3.2. Linear Regression Models

Finally, we create linear regression models using the '`lm`' linear regression model function in R to determine the coefficients of the different textual features using our Sesotho training data and the readability levels computed on the English texts. The structures of the Sesotho linear regression models mimic that of the English readability measures. In this way, we try to ensure that the readability values computed using a particular readability measure are used to create a Sesotho readability measure that uses a similar structure and the same textual features as the English measure.

We then created linear regression models for the different measures. The formulas are presented in Table 4. Our proposed readability formulas for Sesotho maintain a degree of structural preservation for the DCI, CLI, SMOG, FRE, and FKGL formulas. Note that a more simplified version of the CLI formula would use the actual counts and not percentages and result in $CLI_{Sesotho} = -3.683470 + 3.8782(\frac{\#letters}{\#words}) - 72.7569(\frac{\#sentences}{\#words})$.

When comparing the weights of the Sesotho formulas with those of the English formulas, we observe several things. First, there is a reduction in the coefficients of syllables per word within the Sesotho formulas concerning the English ones. For example, this manifests as a heightened and negative weighting for syllables per word within the Sesotho FRE formula.

Second, we propose two structures for the GFI formula. Both versions introduce an intercept for the formula, involving a deduction of 0.177916, which is different from the original formulation. The first proposed formula, $GFI(1)_{Sesotho}$, follows the structure of the original English formula more closely although an intercept is added. The second formula, $GFI(1)_{Sesotho}$ introduces a coefficient to the percentage of complex words, thereby deviating from the original structure.

The English LIX and RIX, do not include weights. To align the readability values that were acquired through the application of English readability measures on the translated examination texts, with the text features observed in the Sesotho texts, it was necessary to introduce weighting factors. This adjustment ensured a more accurate correspondence between the readability values and the adapted for-

	Label	F1	F2	F3	F4	F5	F6	F7	F8
KFGL	F1	1.00	-.96	.93	.87	-.04	.01	.33	.16
FRE	F2	-.96	1.00	-.86	-.85	.08	.01	-.45	-.16
GFI	F3	.93	-.86	1.00	.97	-.02	.03	.29	.18
SMOG	F4	.87	-.85	.97	1.00	-.03	.03	.36	.19
syllables	F5	-.04	.08	-.02	-.03	1.00	.99	-.15	.90
3+ syllables	F6	.01	.01	.03	.03	.99	1.00	.00	.92
%3+ syllables	F7	.33	-.45	.29	.36	-.15	.00	1.00	.14
4+syllables	F8	.16	-.16	.18	.19	.90	.92	.14	1.00

Table 3: The correlation of syllable-based measures and syllable information computed on the Sesotho texts.

Measure	Formula
$FKGL_{Sesotho}$	$= -14.08905 + 0.43405(\frac{\#words}{\#sentences}) + 5.86314(\frac{\#syllables}{\#words})$
$FRE_{Sesotho}$	$= 209.3286 - 1.7930(\frac{\#words}{\#sentences}) - 46.6548(\frac{\#syllables}{\#words})$
$SMOG_{Sesotho}$	$= 0.28788 + 0.68741(\sqrt{\#polysyllabic - words * (\frac{30}{\#sentences})})$
$GFI(1)_{Sesotho}$	$= -4.30942 + 0.28610(\frac{\#words}{\#sentences}) + (\frac{\#complex-words}{\#words})$
$GFI(2)_{Sesotho}$	$= -1.77916 + 0.40861((\frac{\#words}{\#sentences}) + 30.9982(\frac{\#complex-words}{\#words}))$
$CLI_{Sesotho}$	$= -3.683470 + 0.038782(\frac{\#letters}{\#samples} * 100) - 0.727659(\frac{\#sentences}{\#samples} * 100)$
$ARI_{Sesotho}$	$= -13.66031 + 2.87106(\frac{\#letters}{\#words}) + 0.49323(\frac{\#words}{\#sentences})$
$LIX_{Sesotho}$	$= 0.46038 + 1.14736(\frac{\#words}{\#sentences}) + 0.60841(\frac{\#long-words}{\#words} * 100))$
$RIX_{Sesotho}$	$= 0.02180 + 0.76883(\frac{\#long-words}{\#sentences})$
$DCI_{Sesotho}$	$= 4.66547 + 0.14199(\frac{\#words}{\#sentences}) + 0.03264(\frac{\#difficult-words}{\#words} * 100)$

Table 4: Readability measures and corresponding adapted Sesotho formulas.

mulas. For the LIX, the impact of words per sentence is accorded weight, while the percentage of long words remains unaltered. However, in the RIX formula, we ascribe weight to the fraction of long words per sentence. Note that we also introduce intercepts for both the $LIX_{Sesotho}$ and $RIX_{Sesotho}$.

Moreover, a noteworthy decrease in the weight attributed to sentences per word[1] is evident in the Sesotho version of the CLI, when contrasting with its English counterpart. Similarly, the intercept of the $CLI_{Sesotho}$ is appreciably lower compared to the English variant. Similarly, a contrast is discernible in the intercept of the $ARI_{Sesotho}$ formula. Despite the consistent coefficient of words per sentence, the Sesotho ARI entails a reduced weighting of letters per word.

Finally, the coefficient of difficult words appears somewhat lower in the Sesotho CLI formula, as opposed to the English formula. Conversely, the Sesotho CLI formula bestows a higher coefficient for words per sentence.

4. Evaluation

The linear regression summary output provides five statistics to assess the performance of each model and the significance of their coefficients. We consider the Adjusted R-squared, F-statistic, and residual standard error. The outcome of the evaluations is presented in Table 5.

[1]The ratio of the number of sentences to the number of words

	FKGL	FRE	SMOG	GFI1	GFI2	CLI	ARI	LIX	RIX	DCI
F-statistic	293.3	118.3	113.0	124.7	109.1	117.4	433.4	144.0	269.5	23.3
Adjusted R^2	.881	.748	.586	.610	.732	.746	.916	.784	.773	.361
Residual std. error	0.647	4.806	0.873	1.166	0.966	0.848	0.626	2.758	0.441	0.631

Table 5: Evaluations of the adapted linear regression models for Sesotho. Note that the p-values are all significant at $p < .001$.

First, the F-statistic is an indicator of the comprehensive validity of the models. It highlights its statistical significance across all models, as evidenced by the observed p-values ($p < .001$). This affirmation attests to the composite contribution of the predictor attributes in elucidating the variation in text readability, thereby proving that results are highly unlikely to be the result of random chance.

Second, the Adjusted R-squared metric indicates how much the independent variables describe the variance of the data. Our analysis reveals higher values particularly for the $ARI_{Sesotho}$, signifying that the variables of letters per word and sentences per word account for approximately 91% of the predictive capacity associated with ARI scores. However, contrasting outcomes are observed for the SMOG formula, where the number of polysyllabic words accounts for only 59.16% of the overall predictive influence. This variation highlights the varying degrees of contribution made by predictor features across the formulated models.

Finally, the lower residual standard errors describe the standard deviation of the residuals, where lower values indicate better results.

5. Discussion and conclusions

The underlying rationale of this research is the absence of an objective method for identifying the readability levels of texts in Sesotho. We postulate that the already low literacy levels in the indigenous languages of South Africa, including Sesotho, could potentially be worsened by the inappropriate selection of textual materials, especially given the limited pool of texts available in the indigenous languages. We expect that being able to gauge the extent of text readability in the language will assist both learners and teachers in the identification of correctly levelled reading materials. In turn, the use of an objective readability assessment framework will improve access to quality (reading education and hence general) education in Sesotho. However, we acknowledge that the limitations previously ascribed to traditional readability measures remain applicable, even in the context of our proposed adaptations.

Given that Sesotho, like most other indigenous languages in Southern Africa, is a low-resource language, no suitable readability labelled corpora exist. To resolve this issue, educational texts originally written in Sesotho were translated into English. The English translations then formed the source of the readability assessment as traditional English readability measures can be applied. Given the extracted textual features from the Sesotho text, combined with the English readability values, linear regression models can be created. The structure of the linear regression models (e.g., the textual features and how they fit together in the formula) is taken from the corresponding English metrics that were used to compute the readability values.

Among the readability formulas adapted within this article, six depend on the sentence length variable. The adapted Sesotho formulas consistently ascribe greater weight to sentence length in comparison to the original English formulas we adapted to Sesotho. Note that, despite this emphasis, no strong correlations emerge between sentence length and the other variables considered in the sentence length-focused formulas. Nonetheless, the CLI formula's coefficient analysis highlights the impact of a strong negative correlation between sentences per word and words per sentence. This observation accentuates that while sentence length is ascribed higher coefficients in numerous Sesotho formulas, the sentences per word variable receive substantially lower weight in the $CLI_{Sesotho}$ formula, thus underscoring the prominence of the sentence length feature in determining Sesotho text readability. It is, however, also important to note that these features are inverse of each other. As such, they are expected to affect readability levels in contrasting ways.

Furthermore, an examination of the correlation between sentence length and word length in terms of syllable counts reveals a modest negative association. This suggests that as sentence length extends, syllables per word exhibit a slight reduction. In essence, an increase in sentence length corresponds to a marginal decrease in both syllables and letters per word, due to the prevailing negative correlation with letters per word. This observation accentuates sentence length as a dominant predictor of Sesotho text readability.

To the best of our knowledge, the findings of this article present the first formulas for an indigenous language of South Africa and the first for the Sotho-Tswana language group in Southern Africa.

Although our models are trained on educational texts, specifically reading comprehension and summary writing texts, the availability of standardised and objective readability formulas provides a solid starting point for employing machine learning approaches to measure text readability in Sesotho. Furthermore, the methods outlined in this article can be employed in the development of readability measures for other low-resource languages.

5.1. Limitations

Our approach in this article is limited by the reliance on written texts and the existing readability measures that were not originally developed for Sesotho. First, we make the assumption that the translated texts have similar readability. The texts are automatically translated and manually corrected to ensure the most similar texts in English compared to the Sesotho texts. A possible solution would be to develop a text collection that is specifically targeted to readability measures based on Sesotho texts (only). Given that not many texts are publicly available in Sesotho, this will remain a challenge.

Second, to ensure the practical usability of the metrics, an empirical examination involving human participants should be undertaken. Such an evaluation would involve selecting and grading texts using our proposed formulas, and subsequently administering these texts to learners within the grade levels indicated by our formulas. This approach is crucial for evaluating the societal impact of our formulas. However, it also presents a challenge in the need for well-defined criteria to distinguish success or failure in the reading tests that would be administered to participants for the evaluation of the readability levels suggested by our formulas for Sesotho.

Finally, to properly measure the impact of the readability metrics through the effectiveness of the selection of suitable texts, criteria for identifying "success" in reading Sesotho texts will need to be developed. This step is important in the context of Sesotho (and the other South African indigenous languages), in particular given the prevailing challenges that South African learners generally encounter in reading.

5.2. Future studies

The findings discussed in this section reveal a number of avenues for possible future studies. First, a further investigation into the used features may provide more suitable metrics for Sesotho. For instance, the optimal number of letters per word that correspond to long Sesotho words remains an intriguing avenue. This entails scrutinising correlations between differing letter counts per word and the resultant scores to ascertain the highest positively correlated number of counts to the readability levels for defining long words for Sesotho. In this article, we use the English definition for long words.

Second, the utilisation of existing grade levels, albeit untested within the South African education context, underscores an avenue for future research. Future inquiries should investigate the applicability of the FRE, CLI, LIX, and other indicators to South African grade levels to refine the contextual relevance of these measures. Currently, we rely on the existing adaptation of readability scores for the FRE, LIX, and RIX measures within the South African context based on the works of Bargate (2012), and Leopeng (2019). Perhaps future works can consider recalibrating such scores to the South African grades through human-based evaluation methods.

Finally, the metrics are not developed in isolation. The practical use of the metrics will need to be investigated in a proper educational context. Do the metrics indeed allow for the identification of suitable texts for a learner? Can we rely on teachers to evaluate this or do we need other evaluation methodologies? Of course, additionally, we will need to investigate how readers experience the readability metric results. Ultimately, we hope that this research help in improving the reading skills of learners in South Africa which we hope to see in future PIRLS results.

6. Bibliographical References

Karen Bargate. 2012. The readability of managerial accounting and financial management textbooks. *Meditari Accountancy Research*, 20(1):4–20.

Carl-Hugo Björnsson. 1968. *Läsbarhet*. Bokförlaget Liber, Stockholm, Sweden.

Carl-Hugo Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.

Daniel Blanchard. 2011. Unsupervised word segmentation: An investigation of sub-word features. Online on github. https://dan-blanchard.github.io/papers/proposal.pdf.

Vladimir V Bochkarev, Anna V Shevlyakova, and Valery D Solovyev. 2015. The average word length dynamics as an indicator of cultural changes in society. *Social Evolution and History*, 14(2):153–175.

David Carel. 2019. 4 ways to stay safe online-developing a text difficulty indicator for isiXhosa early grades. In *Policy Commons*. Research on Socio-Economic Policy. Available at: https://policycommons.net/artifacts/2105074/

4-ways-to-stay-safe-online/2860372/.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, 165(2):97–135.

Edward Dale and J S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, 20(3):293–325.

Mahmoud El-Haj and Paul Rayson. 2016. Osman: A novel Arabic readability metric. *Procedia Computer Science*, 142:38–49.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221–233.

Rudolph Flesch. 1974. *The art of readable writing*, 2nd edition. Harper, New York.

Thomas François and Cedrick Fairon. 2012. An "AI readability" formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

Government of Lesotho. 1993. *The Constitution of Lesotho*. Government Printer, Maseru.

Robert Gunning. 1952. *The technique of clear writing*. McGraw-Hill: New York.

Robert Gunning. 1969. The fog index after twenty years. *Journal of Business Communication*, 6(2):3–13.

Esté Hefer. 2013. Reading first and second language subtitles: Sesotho viewers reading in Sesotho and English. *Southern African Linguistics and Applied Language Studies*, 31(3):359–373.

Sukhpuneet Kaur, Kulwant Kaur, and Parminder Kaur. 2018. The influence of text statistics and readability indices on measuring university websites. *International Journal of Advanced Research in Computer Science*, 9(1):403–414.

Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and Flesch Reading Ease formula) for navy enlisted personnel. Report, Defense Technical Information Center.

Daleen Krige and Marianne Reid. 2017. A pilot investigation into the readability of Sesotho health information pamphlets. *Communitas*, 22:113–123.

Nancy C Kula and Lutz Marten. 2008. Central, east and southern african languages. In Peter Austin, editor, *One Thousand Languages*, pages 86–111. Ivy Press/University of California Press.

Sanja Kusec, Miroslav Mastilica, Gordana Pavlekovic, and Luka Kovacic. 2002. Readability of patient information on diabetes on the Croatian Web sites. In *Health Data in the Information Society, Netherlands*, pages 128–132. IOS Press, Amsterdam.

Sandra Land. 2015. Reading and the orthography of isiZulu. *South African Journal of African Languages*, 35(2):163–175.

Makiti Thelma Leopeng. 2019. *Translations of informed consent documents for clinical trials in South Africa: Are they readable?* Thesis, University of Cape Town.

Georgia Loukatou. 2019. From phonemes to morphemes: Relating linguistic complexity to unsupervised word over-segmentation. In *Proceedings of TyP-NLP: The First Workshop on Typology for Polyglot NLP, Florence, Italy*, New Jersey, USA. Association of Computational Linguistics.

Tshokolo J Makutoane. 2022. 'The people divided by a common language': The orthography of Sesotho in Lesotho, South Africa, and the implications for Bible translation. *HTS Teologiese Studies/Theological Studies*, 78(1):9.

Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Hans Messerschmidt, JJE Messerschmidt, and DP Thulo. 2003. A human-assisted computer generated LA-grammar for simple sentences in Southern Sotho. *Southern African Linguistics and Applied Language Studies*, 21(1–2):41–47.

Lubna Ali Mohammed, Musheer Abdulwahid Aljaberi, Antony Sheela Anmary, and Mohammed Abdulkhaleq. 2023. Analysing English for science and technology reading texts using Flesch Reading Ease online formula: the preparation

for academic reading. In *International Conference on Emerging Technologies and Intelligent Systems*, pages 546–561, New York. Springer.

Lehlohonolo Mohasi and Daniel Mashao. 2005. Phonetization for text-to-speech synthesis in Sesotho. In *The sixteenth annual symposium of the Pattern Recognition Association of South Africa*, pages 121–122, Langebaan, South Africa. Citeseer.

Rosalia Moroesi Nakin. 2009. *An examination of language planning and policy in the Eastern Cape with specific reference to Sesotho: A socioliguistic study*. Ph.D. thesis, Nelson Mandela Metropolitan University, South Africa.

Moselane Andrew Nhlapo. 2021. *Historical perspectives on the development of Sesotho linguistics with reference to syntactic categories*. Ph.D. thesis, University of the Free State, South Africa.

Parliament of Zimbabwe. 2021. *The Constitution of Zimbabwe*. Veritas, Harare.

Michelle Rasheed. 2023. Kindling a desire to read: A review of three young adult novels. *South Carolina Association for Middle Level Education Journal*, 2(1):109–113.

Marianne Reid, Mariette Nel, and Ega Janse Van Rensburg-Bonthuyzen. 2019. Development of a Sesotho health literacy test in a South African context. *African journal of primary health care & family medicine*, 11(1):1–13.

Republic of South Africa. 2023. *Constitution eighteenth amendment bill*. Department of Justice and Correctional Services, Pretoria.

Karen Roux, S van Staden, and M Tshele. 2021. *Progress in International Reading Literacy Study 2021: South African Preliminary Highlights Report*. Department of Basic Education, Pretoria, South Africa.

Ntaoleng Belina Sekere. 2004. *Sociolinguistic variation in spoken and written Sesotho: A case study of speech varieties in Qwaqwa*. Thesis, University of South Africa.

Mmasibidi Setaka. 2018. *Corpus-based Lexicography for Sesotho*. Ph.D. thesis, University of Pretoria, South Africa.

Mmasibidi Setaka and Danie J Prinsloo. 2020. A critical evaluation of three sesotho dictionaries. *Lexikos*, 30:445–469.

Johannes Sibeko. 2022. Tshebediso ya melao kabong ya dinoko tsa Sesotho. *Southern African Linguistics and Applied Language Studies*, 40(4):494–506.

Johannes Sibeko. 2023. Using classical readability formulas to measure text readability in Sesotho. In Tomaž Erjavec and Maria Eskevich, editors, *Selected papers from the CLARIN Annual Conference 2022*, volume 198, pages 120—132. Linköping Electronic Conference Proceedings, Prague, Czechia.

Johannes Sibeko. 2024. Harnessing google translations to develop a readability corpus for sesotho: An exploratory study. *Journal of the Digital Humanities Association of Southern Africa*, 5:1–12.

Johannes Sibeko and Orphée De Clercq. 2023. A corpus-based list of frequently used words in Sesotho. In *Proceedings of the Fourth workshop on Resources for African Indigenous Language (RAIL 2023), Dubrovnik, Croatia*, pages 32–41, New Brunswick, New Jersey, USA. Association for Computational Linguistics.

Johannes Sibeko and Menno Van Zaanen. 2021. An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, 3(1):1–11.

Johannes Sibeko and Menno Van Zaanen. 2023. A data set of final year high school examination texts of South African home and first additional language subjects. *Journal of Open Humanities Data*, 9(9):1–6.

Edgar A Smith and R.J Senter. 1967. *Automated readability index*. Clearing house for Federal Scientific and Technical information, Cincinnati, Ohio.

Elsje van Bergen, Margaret J Snowling, Eveline L de Zeeuw, Catharina EM van Beijsterveldt, Conor V Dolan, and Dorret I Boomsma. 2018. Why do children read more? the influence of reading ability on voluntary reading practices. *Journal of Child Psychology and Psychiatry*, 59(11):1205–1214.

Phillip Van Oosten, Dries Tanghe, and Véronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC 2010), Valletta, Malta*, pages 775–782, Paris. European Language Resources Association (ELRA).

Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory and Practice in Language Studies*, 2(1):43–53.

Shixiang Zhou, Heejin Jeong, and Paul A Green. 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication*, 6:97–111.

Johannes C Ziegler, Conrad Perry, Arthur M Jacobs, and Mario Braun. 2001. Identical words are read differently in different languages. *Psychological science*, 12(5):379–384.

Bootstrapping syntactic resources from isiZulu to Siswati

Laurette Marais[1], **Laurette Pretorius**[2], **Lionel Posthumus**[3]
[1]Voice Computing Research group, CSIR
[2]Division of Computer Science, Department of Mathematical Sciences, Stellenbosch University
[3]CALT@UJ, University of Johannesburg
lmarais@csir.co.za, lpretorius@sun.ac.za, lionelp@uj.ac.za

Abstract

IsiZulu and Siswati are mutually intelligible languages that are considered under-resourced despite their status as official languages. Even so, the available digital and computational language resources for isiZulu significantly outstrip those for Siswati, such that it is worth investigating to what degree bootstrapping approaches can be leveraged to develop resources for Siswati. In this paper, we present the development of a computational grammar and parallel treebank, based on parallel linguistic descriptions of the two languages.

Keywords: Grammatical Framework, parallel treebanks, computational grammar

1. Introduction

IsiZulu and Siswati[1] are Southern Bantu languages that belong to the Nguni group, and as such are morphologically rich languages that have a noun class system which in turn generates concordial agreement. The Nguni languages have a conjunctive orthography and also exhibit significant morphophonological affixing, leading to long tokens for which morphological analysis is non-trivial.

The Nguni languages are mutually intelligible (Ndhlovu, 2022), and this characteristic allows for exploitation in an under-resourced context. While isiZulu is an official language of South Africa[2], and Siswati an official language of South Africa and the Kingdom of Eswatini[3], they are both under-resourced, Siswati significantly more so than isiZulu (Moors et al., 2018).

Previous work by Bosch et al. (2008) showed the feasibility of bootstrapping finite state morphological analysers following a systematic approach. In this case, isiZulu served as the starting point from which resources for other Nguni languages could be developed. Some of the key findings of this work was that bootstrapping between the Nguni languages drastically reduces development time, which can be significant in the context of under-resourced languages. A bootstrapping approach also results in special focus being given to the differences between the languages: "By exploiting correspondences and linguistic relatedness, more effort may be spent on those aspects in which the languages differ, ensuring end products of superior quality, both linguistically and computationally." (Bosch et al., 2008, p. 85)

A natural next step would be to explore application of the bootstrapping approach beyond morphology to syntax. Our point of departure for this work is the Grammatical Framework (GF) isiZulu resource grammar, with the primary goal of bootstrapping a Siswati resource grammar. In the process, we develop a parallel treebank by hand, which we then augment using the parallel resource grammars to achieve a larger semi-synthetic treebank - a first for Siswati. We evaluate the resource grammars by manually evaluating a subset of the augmented data to ensure that the functions of the grammars behave as expected when combined in new ways.

We based our bootstrapping methodology on a set of two textbooks, on isiZulu and Siswati respectively, in order to ensure a systematic and linguistically aware approach. Even here, the Siswati textbook (Taljaard et al., 1991) is "largely based on" the isiZulu textbook (Taljaard and Bosch, 1988) and features the two authors of the isiZulu book alongside a specialist Siswati linguist. In a certain sense, we rely on the "bootstrapping" of high quality linguistic descriptions of the language by linguists in order to guide a systematic and reliable bootstrapping approach to computational resources.

2. Background

Bootstrapping of resource grammars, specifically GF resource grammars, has been done for various related languages, with the most relevant being the work on Runyankore and Rukiga by Nabende et al. (2020), as well as the work on the Kenyan Bantu Languages (Ekegusii, Kikamba and Swahili) by Kituku et al. (2021). Due to the under-resourced status of these languages, suitable evaluation corpora do not exist and require special development. Con-

[1]The three letter language codes for isiZulu and Siswati are zul and ssw respectively.

[2]IsiZulu has the largest number of L1 speakers of all the (Nguni) languages, namely around 15 million, while Siswati has around 3 million.

[3]Also known by its former official name Swaziland.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 77–85

sequently, a full evaluation of the resource grammars for Runyankore and Rukiga has not been reported on. Evaluation for the Kenyan Bantu languages was focused on software engineering aspects of bootstrapping, with no specific mention of the final correctness of the grammars. The language fragments used in iterative testing during development were translated from English examples illustrating the purpose of each function in the grammar. In terms of coverage, then Kenyan Bantu language resource grammars are not as mature as the isiZulu resource grammar. For example, in the GF Github repository, the Kenyan Bantu language functor only contains one function for constructing verb phrases, namely `UseV`, which is used for intransitive verbs. The isiZulu (and now Siswati) resource grammars, by contrast, include 21 functions for constructing verb phrases, covering also transitive verbs, the reflexive construction, the copulative constructions, adverbial modification and verbs with verb and sentence complements.

Therefore, although previous GF work exists for other Bantu languages, it is difficult to provide a direct comparison of our work to these other efforts.

Our aim is to exploit existing linguistic resources for isiZulu and Siswati in order to base our bootstrapping of the Siswati resource grammar on a systematic and parallel exposition of the linguistic characteristics of the two languages.

3. Comparison of isiZulu and Siswati

As in all Bantu languages, the structure of isiZulu and Siswati is based on two principles, viz. nominal classification (the system of noun classes) and concordial agreement across various word categories (the system of concords). (These are but 2 outstanding characteristics of the Bantu languages.)

Generally speaking, the noun consists of two main parts, viz. a noun class prefix and a noun root/stem. Furthermore, every noun belongs to a so-called noun class by virtue of the form of its prefix, also referred to as its class gender. This notion of class gender is significant since it generates grammatical agreement by means of these class prefixes, also termed gender number prefixes. These noun classes are numbered, with the noun class system of isiZulu and Siswati being very similar.

A concord is a structural element (agreement marker/morpheme) which formally marks the relationship between a noun and all other words in a sentence that have a direct semantic-syntactic relationship with the noun. The above-mentioned gender agreement must be observed in all parts of the utterance which are directly linked to the noun. Therefore, we say that word categories such as verbs, pronouns, adjectives, relatives, possessives etc. are brought into concordial (i.e. grammatical) agreement by means of these concords. Examples (1) and (2) show an isiZulu and a Siswati sentence, respectively.

(1) *Leli bhubesi li-zo-yi-luma*
Dem5 NStem5 SC5-Fut-OC9-VStem
in-komo ya-mi
NStem9 PC9-PPron1PSg
'This lion will bite my cow.'

(2) *Leli-bhubesi li-to-yi-luma*
Dem5-NStem5 SC5-Fut-OC9-VStem
in-khomo ya-mi
NStem9 PC9-PPronP1Sg
'This lion will bite my cow.'

Before listing a number of systematic differences between isiZulu and Siswati that we exploit in our bootstrapping process, we take a closer look at examples (1) and (2). The one noun root *-bhubesi*, the verb stem, the class 5 demonstrative, the class 5 subject concord, the class 9 object concord, the class 9 possessive concord and the possessive pronoun, first person singular, are identical. Moreover, in both languages the noun root for 'cow' is *-khomo*. However, in isiZulu the class 9 surface form is subject to a morphophonological alternation rule and is realised as *-komo*. Finally, the future morpheme is *-zo* in isiZulu and *-to* in Siswati.

As a point of departure, important regular morphophonological differences between the two languages may be systematised as follows (Mordaunt et al., 2023; Bosch et al., 2008; Taljaard and Bosch, 1988; Taljaard et al., 1991):

1. The alphabet and click omission: While both languages use the Latin alphabet (A-Z), Siswati omits Q and X, while in isiZulu /q/ and /x/ represent click consonants. In isiZulu the click sounds /c/, /q/ and /x/ are represented by the click sound /c/ in Siswati. for example, *-qina* (zul) and *-cina* (ssw) both mean 'be hard'.

2. Consonant substitution or addition: The /z/ that often occurs in isiZulu roots/stems and in the class 8 and 10 prefixes and concords, is usually substituted with /t/ in Siswati. for example, *-zama* (zul) and *-tama* (ssw) both mean 'try'.

   The /th/ and /t/ in isiZulu is usually realised as /tf/ when followed by /o/, /u/ and /w/, and as /ts/ when followed by /a/, /e/ and /i/ in Siswati. Examples are *-thola* (zul) and *-tfola* (ssw), which mean 'find', and *-thatha* (zul) and *-tsatsa* (ssw), which mean 'take'.

   The /d/ in isiZulu converts to /dv/ when followed by /o/, /u/ and /w/, and to /dz/ when followed by /a/, /e/ and /i/ in Siswati, for example *-dubula* (zul) and *-dvubula* (ssw), meaning

'shoot', and *-dabula* (zul) and *-dzabula* (ssw), meaning 'tear'.

Other differences are the consonant clusters /mp/ and /nk/ in isiZulu that become /mph/ and /nkh/ in Siswati, for example *impendulo* (zul) and *imphendulo* (ssw), meaning 'reply'.

3. Pre-prefix vowel deletion, addition and substitution: The isiZulu noun class prefix consists of a consonant-vowel sequence (also referred to as the basic prefix), preceded by a so-called augment (also referred to as class pre-prefix), a preceding copy vowel that fulfils different grammatical functions, e.g. definiteness and specificity, and is subject to morphophonological processes such as vowel deletion and coalescence. In Siswati this augment is only present in classes 1, 3, 4 and 6 and in class 9 (where it precedes a nasal consonant). Moreover, in class 6 this pre-prefix is /e/ and not /a/. An example is ***a****makati* (zul) and ***e****makati* (ssw) for 'cats'.

4. The relative construction and concords: Whereas the relative construction in isiZulu has *a-* as so-called relative morpheme, the relative morpheme in Siswati is *la-*. In both languages the *a-* and *la-* respectively assimilates with the vowel of the basic prefix and vowel coalescence takes place across the consonant to form the relative concord. An example is *umfana omunye* (zul) and *umfana lomunye* (ssw), from *a+munye* and *la+munye*, meaning 'another boy'.

5. Lexical items: While the two languages share many noun and verb roots/stems, lexically there are differences, for example *-phuza* (zul) and *-natsa* (ssw), meaning 'drink'.

6. Orthography: In isiZulu, demonstratives are written disjunctively from the noun that follows, while in Siswati the first position demonstrative ('this/these') is written conjunctively with the following noun, as in example (1): *leli bhubesi* (zul) versus *lelibhubesi* (ssw), meaning 'this lion'.

7. The imperative: In isiZulu monosyllabic verb stems, *yi-* or *i-* are prefixed or *-na* suffixed to the stem for the imperative directed at one person. In Siswati *-ni* is suffixed to the verb stem, for example ***Yi****dla/****I****dla/Dla****na****!* (zul) and *Dla****ni****!* (ssw), meaning 'Eat!', directed to one person.

In summary, the differences 1-3 above apply to the two languages across all constructions and lexical items. Complementary to this general exposition, are the word category and grammatical construction based parallel expositions of Taljaard and Bosch (1988) and Taljaard et al. (1991), the latter two providing practical grammar orientated perspectives, ideally suitable for direct application to and implementation in the bootstrapping of the Siswati grammar from the isiZulu RG.

4. GF isiZulu resource grammar

The isiZulu resource grammar (isiZulu RG) used in this work is implemented in Grammatical Framework (GF), a computational grammar framework for the development of multilingual grammars. The framework utilises an interlingua architecture, such that a GF grammar consists of an abstract syntax and one or more concrete syntaxes, one for each language. Abstract categories and functions are defined in the abstract syntax, which are implemented in the concrete syntaxes as linearisation categories and linearisation functions. The GF runtime enables linearisation of abstract syntax trees into natural language strings, as well as parsing of natural language strings into abstract syntax trees (Ranta, 2011).

GF resource grammars typically form part of the Resource Grammar Library (RGL), which shares a common abstract syntax and custom extensions between over 40 languages (Ranta et al., 2020). The categories and functions are syntactic in nature, with categories for nouns, noun phrases, verbs, verb phases, adverbial phrases, clauses, sentences, etc., along with functions for combining these categories into tree structures.

Originally, the intent of the RGL was to serve as a linguistic software library to enable rapid development of application specific grammars (Ranta, 2009). The implementation of the syntactic categories and functions would capture the general morphology and syntax of the language, which could then be reused by application grammars for specific use cases. More recently, however, attempts have been made to employ the general use grammars of the RGL towards wide-coverage parsing as well as for bootstrapping Universal Dependencies treebanks (Ranta et al., 2020).

The isiZulu RG models the morphology and syntax of isiZulu via the implementation of some functions from the RGL common abstract syntax, in addition to a set of extra language specific abstract functions (Marais and Pretorius, 2023b).[4]

Following an approach typical for the implementation of Bantu languages, the isiZulu RG models the language at the subword level. In short, this means that the base tokens of the grammar do

[4]See the README at https://github.com/GrammaticalFramework/gf-rgl/blob/master/src/zulu/README.md

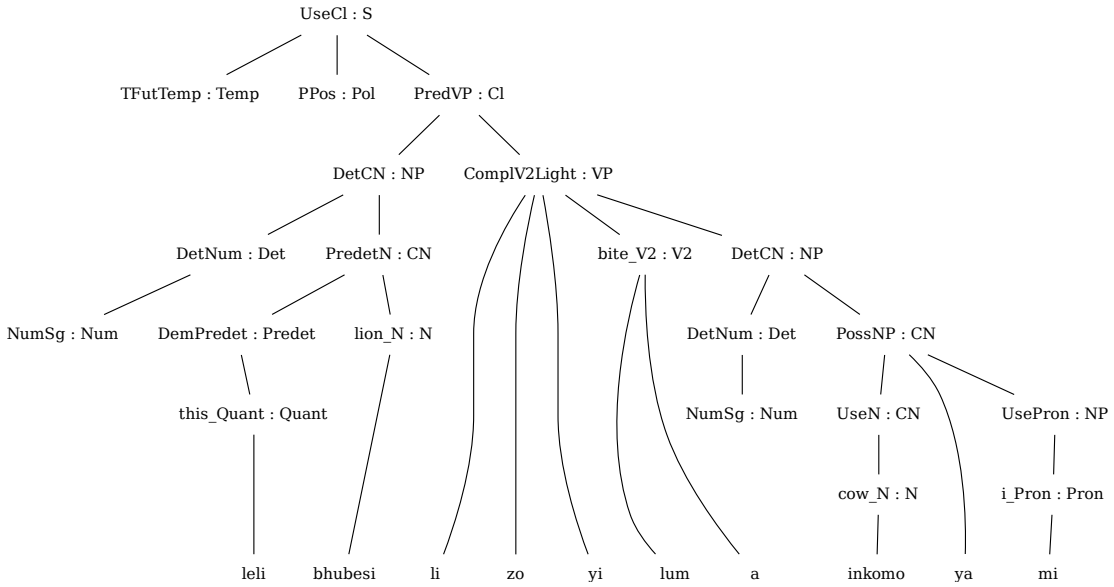


Figure 1: GF parse tree for example (1)

not correspond to orthographic words but to subword segments, which are glued together at runtime using built-in orthography engineering support in the GF C-runtime (Angelov, 2015). An example of this is given in Figure 1, showing how the surface segments of the isiZulu sentence in example (1) in Section 3 are produced by different functions in the isiZulu RG. We will say more about how morphophonological alternation is modelled in Section 6.3.

5. Methodology

Our methodology is depicted in Figure 2. We started with two resources (shown in blue) and from them developed three new resources (shown in orange). The isiZulu RG forms the computational basis for the work, with the set of parallel textbooks providing the linguistic information required to develop and evaluate a new Siswati resource grammar.

The isiZulu RG has so far been used to expand morphosyntatically complex entries in the isiZulu Wordnet (Marais and Pretorius, 2023a), as the general purpose syntactic parser for isiZulu (Marais and Pretorius, 2023b) and as a mechanism for generating annotated data for training morphological segmentation models for isiZulu (Mkhwanazi and Marais, 2024). We therefore consider it to be a mature model of isiZulu and a suitable basis upon which to develop similar models for related languages.

The parallel texts provide us with two kinds of information, namely a parallel linguistic exposition of the two languages, as well as high quality parallel example sentences exhibiting the linguistic features described in the books. The parallel linguistic exposition served as the basis for the development of the Siswati RG, while the parallel examples were used to create a parallel development treebank. Here, the isiZulu RG was used to parse the examples to speed up the process of obtaining a tree representation for each parallel sentence pair.

The treebank itself served as a regression test during development to ensure that adaptations for the Siswati rendered the correct linearisations (natural language strings) from the trees, and it also served to ensure that no errors were introduced in the process of some superficial refactoring of the isiZulu RG in order to minimise code divergence. We give more detail about this process in Section 6.

The final evaluation involved the creation of an augmented treebank based on the one used in development. It was created using a few basic rules defining tree modifications and applied to the development treebank. From the newly created trees, linearisations in both isiZulu and Siswati were generated, and these were manually evaluated. This would ensure that the adaptations that were made to the Siswati on the basis of the linguistic exposition and evaluated during development on the parallel treebank, would generalise to new trees.

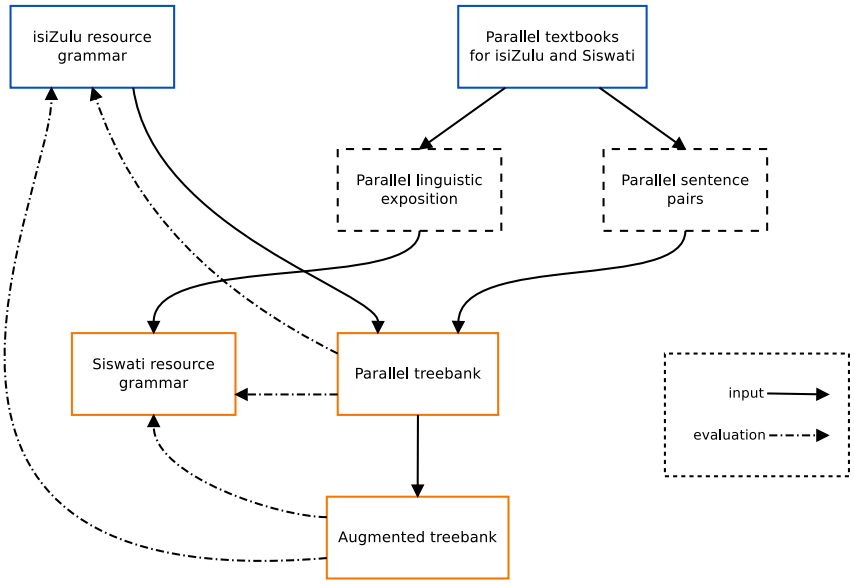


Figure 2: Methodology

6. Adapting the isiZulu grammar to Siswati

6.1. Software logistics

The most naïve way to bootstrap a new GF RG from an existing one is to make a copy of all the relevant files and to change them in specific ways. The GF requirement is that concrete module names have the form 'XY.gf', where X is the name of the abstract module being implement and Y is a code indicating the specific language, ideally based on the language's 3-letter ISO code. The headers of concrete modules also contain this code, and hence a first step would be to systematically change the file and header names. The new resource grammar therefore starts out as an exact copy of the original, which can be changed in precisely those places where the two languages differ.

Of course, the original resource grammar may also be added to or changed while work is ongoing on the new grammar, which could soon cause unnecessary code divergence. GF encourages modular design of grammars, along with the use of functors to model closely related languages according to sound software engineering principles (Ranta, 2009). However, the right moment to functorise a parallel implementation depends on having a good understanding of how the similarities and differences between two or more languages should be modelled. Our intent is to extend this bootstrapping approach to the other Nguni languages and beyond, and hence we have opted not to implement a functor yet, since a prematurely implemented one could turn out to be more of a hindrance than a help.

In order to minimise code divergence following a purely parallel implementation approach, the module system of GF was exploited so that all strings in the grammar (apart from roots and stems included in lexicon modules) are contained in the two main resource modules, namely ResZul.gf and ResSsw.gf. These strings could then be accessed by other modules exclusively via operations defined in the resource modules. This would ensure that differences at the orthographic level would be defined entirely in the respective resource modules, while morphosyntactic differences would be defined in the relevant linearisation functions of the concrete modules. The differences can therefore be monitored at a glance using software that indicates line differences between files, such as `diff`.

The isiZulu RG included a number of custom abstract modules, modelling aspects of isiZulu not found in the common abstract syntax. These files were moved to a folder named 'nguni', so that they could be utilised by both resource grammars.

6.2. Linguistically-driven adaptation

The parallel linguistic exposition of the two textbooks provided a practical and systematic basis for adapting the Siswati RG from the isiZulu. In contrast to the description of similarities and differences as summarised in Section 3, the parallel texts provided a map of where these differences manifest in the respective languages, which simplified the process of identifying which functions and operations in the resource grammars would be different. Not all constructions in the textbook are implemented in the isiZulu RG: we limited the scope of the adaptation to what is currently stable in the isiZulu RG, having established that it has already been used in a number of applications. As such, we excluded from this adaptation the situative

mood, certain interrogative constructions, auxiliary verbs and indirect relatives.

From Section 3 it is clear that morphophonology would play a central role in any bootstrapping effort. We next provide a short description of how morphophonological alternation is modelled in the isiZulu and subsequently the Siswati resource grammars.

6.3. Morphophonological alternation in GF

In GF, morphophonological alternation can be modelled by defining alternative forms of certain morphemes and selecting the correct form to use in a specific context based on one or more parameters supplied by the context. This is necessitated by the fact that the strings of a GF grammar cannot be inspected at runtime, only at compile time.

For example, due to morpheme fusion, the form of the possessive concord depends on the initial sound of the noun or pronoun to which it is prefixed. A parameter called `RInit` is used to keep track of this at runtime, defined to distinguish between the different vowels (with values `RA` to `RU`, as shown in Figure 3) and consonants as a whole (with value `RC`). The table containing the possessive concord is essentially 2-dimensional, with the first dimension representing the agreement information of the possessee, while the second dimension represents the initial sound of the possessor noun or pronoun. Figure 3 shows how this in encoded in a GF table.

Agreement is encoded as a compound parameter in which the first value is a constructor dealing with grammatical person, and the subsequent values deal with grammatical number and class gender where applicable. For example, `First Sg` refers to agreement with the first person singular pronoun, while `Third C3_4 Pl` refers to agreement with plural nouns of classes 3 and 4.

A significant number of adaptations to the Siswati resource grammar consisted of systematically altering the strings contained in tables such as these.

6.4. Changes to the Siswati resource module

Recall that the respective resource modules of the resource grammars were designed to contain the majority of differences between the two languages by containing all strings used in the grammar (apart from a lexicon). In this section we discuss changes made to the ResSsw.gf module, unless otherwise indicated.

The centrality of the noun class system makes nouns an obvious place to start, which is most likely why the two textbooks also devote the first few chapters to nouns, their classes and the associated prefixes. This is dealt with in the resource modules of the RGs in two main operations, `nomNoun` and `locNoun`, for nouns and locativised nouns. Supporting operations deal with the morphophonological alternation which occurs when noun roots/stems are joined with the relevant prefixes and suffixes. These were the first adaptations to be made to the Siswati RG.

The focus then shifted to verbs, starting with alternation that occurs within the verb root/stem, especially as it relates to the verb-final morpheme. After that, the various pre-root verbal morphemes were adapted by making changes to the subject and object concord tables, as well as to the operations for producing the appropriate forms of the tense markers and relative prefix. The forms of the reflexive prefix and relative suffix were also changed.

These changes were sufficient to also cover most of the changes necessary for correctly modelling the copulative constructions, although additional changes to the identifying copulative marker and the adjectival concord were also required. In fact, the identifying copulative prefix is not required in the Siswati grammar, which amounted to a syntactic change that was made in the `VerbExt` module. For example, in isiZulu the sentence 'The lion is an animal' is expressed as *Ibhubesi **yi**silwane*, while in Siswati it is expressed as *Libhubesi silwane* (Taljaard and Bosch, 1988; Taljaard et al., 1991).

The tables containing the absolute, possessive and all three sets of demonstrative pronouns were also changed, along with the possessive and quantitative concords.

Finally, the various adverbial prefixes were changed. This was, perhaps surprisingly, one of the more substantial changes required. In isiZulu, the morphophonological alternation of adverbial prefixes like *nga-* and *njenga-* is based on the class prefix of the noun to which it is prefixed, whereas in Siswati, the alternation is based directly on the class to which the noun belongs, regardless of the form of its prefix. The sound changes also follow a different pattern with regards to the classes compared to isiZulu. Hence, instead of altering strings in a table, the structure of the tables in which the adverbial prefixes were housed was changed, accurately reflecting this difference between the languages.

The other syntactically significant changes that were implemented relate to the imperative, since the morphosyntactic structure of imperatives differ between the two languages when it comes to monosyllabic verb stems and the copulative constructions. These changes were implemented in all modules containing functions for constructing `VP`s (verb phrases).

The most important insight gained during the process of bootstrapping from one Nguni language to another is the centrality of a transparent and sys-

```
param RInit = RA | RE | RI | RO | RU | RC ;

oper poss_concord_agr : Agr => RInit => Str = table {
    First Sg => table {(RA|RC) => "wa" ; (RE|RI) => "we" ; (RO|RU) => "wo" } ;
    First Pl => table {(RA|RC) => "ba" ; (RE|RI) => "be" ; (RO|RU) => "bo" } ;
    ...
    Third C3_4 Sg => table {(RA|RC) => "wa" ; (RE|RI) => "we" ; (RO|RU) => "wo" } ;
    Third C3_4 Pl => table {(RA|RC) => "ya" ; (RE|RI) => "ye" ; (RO|RU) => "yo" } ;
    Third C5_6 Sg => table {(RA|RC) => "la" ; (RE|RI) => "le" ; (RO|RU) => "lo" } ;
    Third C5_6 Pl => table {(RA|RC) => "a" ; (RE|RI) => "e" ; (RO|RU) => "o" } ;
    ...
  } ;
```

Figure 3: Table for the possessive concord, parameterised to contain alternative forms based on the initial sound of the possessor

tematic model of morphophonology. This ensured that the majority of changes required related to the strings in the resource modules that represent morphemes alongside their morphophonological alternatives, with very few changes requiring a more substantial structural change.

7. Developing a parallel treebank

Manually capturing parallel sentences from textbooks and obtaining trees to represent them is a time consuming and therefore expensive task. Consequently, we opted to select about four to five structurally dissimilar sentences from each relevant chapter of the parallel textbooks, although the capturing of all the sentences in the textbooks is continuing. In some places, the same linguistic construction was illustrated in the textbooks using multiple sentences with alternative word orders, some of which have not been included in the isiZulu RG. In such cases, we included the sentences whose word order is already implemented in the resource grammar. While it is in principle possible to implement functions for alternative word orders, the decision to do so must also weigh the computational cost associated with a larger grammar and will be considered in future, as well as the expected frequency in which the alternative word order appears in isiZulu and Siswati corpora. Moreover, the purpose of this work was to bootstrap the existing isiZulu grammar, which we consider to be mature. Inclusion of the additional sentences in the treebank, along with the implementation of functions to support them, is considered future work.

7.1. Obtaining trees

The process of finding trees to represent the sentences was somewhat expedited by employing the GF runtime as a parser. IsiZulu sentences were parsed using the isiZulu resource grammar, along with a large isiZulu lexicon. In almost all cases, the correct tree was selected from among those provided by the runtime. In cases where the syntactic ambiguities of the sentence made selecting from a large number of possible parses difficult, the correct tree was developed by hand on the basis of the context within which it is provided in the textbooks, as well as its English gloss. It was then linearised to isiZulu in order to confirm its correctness.

In this way, a tree was found for 125 pairs of sentences, covering the chapters on nouns, concordial agreement in verbs, adverbial forms, the various tenses of the verb, absolute and demonstrative pronouns, copulative forms, direct relatives, the enumerative, numerals, and the subjunctive form. While this would constitute, to our knowledge, the first treebank for Siswati, it is admittedly quite small. However, in stark contrast to one that would be based on a corpus, the treebank was designed specifically to test a wide variety of linguistic constructions and can therefore be said to be highly representative of the languages. For that reason, it is ideal as the basis for continuous evaluation of a computational grammar during development.

7.2. Lexicon support

The trees as they were developed via parsing using the isiZulu RG, which was paired with a large isiZulu lexicon, included lexical functions based on isiZulu roots and stems. For instance, the tree would use the function `theng_V2` for trees in which the verb *-thenga* (to buy) appeared. Since no computational lexicon currently exists for Siswati, the required lexical functions for modelling the Siswati sentences had yet to be developed.

Consequently, a bilingual isiZulu-Siswati lexical database was manually developed from the sentences in the treebank. To improve future interoperability with multilingual systems, entries were given English-based function names. The information necessary to derive parallel concrete GF lexicon modules was added for isiZulu and Siswati, such as the relevant root or stem and class information

for nouns. The lexicon size is 190 functions, of which 76 are for nouns (eg. `student_N`) and 89 are for verbs (eg. `come_V`).

Our focus was to develop and evaluate the morphosyntactic functions for a Siswati resource grammar from the existing isiZulu, for which a limited yet representative lexicon is sufficient. An essential resource that must still be developed is a large Siswati computational lexicon.

8. Evaluation

During development, continuous evaluation relied on the manually developed treebank and hence continued until regression tests for both isiZulu and Siswati succeeded. Now, it was time to evaluate the ability of the grammar to generalise to unseen combinations of functions.

While it is possible to use random generation of trees for evaluation, trees generated in such a way are often nonsensical. This limits the value of having them evaluated, and also confronts the evaluator with a difficult task, especially in case of failure: did the grammar render a meaningful tree incorrectly or did it "correctly" render a nonsensical one? Even the task of determining whether a tree represents a meaningful sentence can be difficult, with different kinds and degrees of problematic combinations of functions possibly occurring. False positives may also undermine the evaluation process.

Instead, in order to test the Siswati RG, an augmentation strategy was defined according to which each tree in the manually developed treebank was modified by randomly selecting from a list of possible modifications. These included swapping tense, polarity, number, subject nouns and pronouns. In this way, the same basic linguistic structures were retained in the new test set, but the syntactic context in which they occurred was changed in a guided yet randomised way.

This led to a new set of 125 trees, each with their isiZulu and Siswati linearisations produced by the respective resource grammars. The linearisations were then manually evaluated and errors categorised. Table 1 gives the outcome of the evaluation. Note that in all cases, errors either occurred in both languages or in none, indicating that the bootstrapping itself was entirely successful, i.e. the small percentage of grammatical errors was carried over from the isiZulu RG.

The first thing to note about the results is that inaccurate augmentation occurred for 14 trees (about 11%), often due to unidiomatic or ungrammatical use of lexical items. Making small changes to trees could place words in a syntactic context that was in some way problematic. This highlights that although this kind of augmentation can be very powerful, care has to be taken when designing tree modification rules to limit their application to appropriate contexts.

In three cases, small inaccuracies in the original parallel treebank, originating from the textbooks, were discovered, which we named seed errors. For both the augmentation and seed errors, the grammar still succeeded in producing reasonable, and in most cases morphologically acceptable, linearisations for problematic trees. The number of true grammar errors amounts to less than 2% of the treebank. This is a very encouraging result.

9. Conclusion

We have presented a bootstrapping process to develop a Siswati GF RG from the existing isiZulu RG. To aid in development and evaluation, a set of parallel textbooks was employed, which had themselves been "bootstrapped" due to the similarity of the languages. The parallel texts provided a practical and systemic basis for implementing known differences between the languages, as well as a set of high quality parallel sentences. These were used to develop manual and augmented parallel treebanks, which were utilised during development and evaluation[5].

Our work confirms the feasibility of such bootstrapping approaches for closely related languages. The isiZulu GF resource grammar was developed over a three-year period[6], while the Siswati resource grammar could be developed and evaluated in less than a year[7]. Such reductions in effort and cost are especially important in resource development for under-resourced languages, since their under-resourced status often relates as much to human and financial resources as to language resources.

We intend to explore a number of avenues for continued work. A refined set of tree modification rules could be utilised to further augment the manually developed parallel treebank, which in turn could be converted to a parallel Universal Dependencies treebank (Kolachina and Ranta, 2019) and used to bootstrap UD parsers for both isiZulu and Siswati. This would require the development of improved lexical resources, especially for Siswati. We may look to exploring the possibility of exploiting known orthographic and phonological differences, as discussed in Section 3, to enable this development from existing isiZulu lexical resources, taking care to deal with lexical differences accurately.

[5] https://github.com/LauretteM/gf-bantu-resources

[6] https://shorturl.at/pyUX3

[7] https://github.com/GrammaticalFramework/gf-rgl

Result	Description	Number
Tree error	The new tree is syntactically problematic	6
Lexical error	The new tree uses a word in the wrong syntactic context	8
Seed error	There was a problem with the original sentence	3
Grammar error	The grammar produced an incorrect linearisation	2
Correct	No problem with the new tree or its linearisations	106

Table 1: Summary of evaluation result on the augmented treebank

We also intend to repeat the bootstrapping process for isiXhosa (a relatively large Nguni language with around 8 million L1 speakers) and isiNdebele (a relatively small Nguni language with around 1 million L1 speakers), incorporating the insights gained from developing the Siswati RG. From there, resource grammars for other Southern Bantu languages beyond the Nguni group could be targeted.

We hope in this way to continue to build upon comparative linguistic research to develop digital language resources for the under-resourced languages of South Africa.

10. Bibliographical References

Krasimir Angelov. 2015. Orthography engineering in Grammatical Framework. In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) 2015 Workshop*, pages 33–40.

Sonja Bosch, Laurette Pretorius, and Axel Fleisch. 2008. Experimental Bootstrapping of Morphological Analysers for Nguni Languages. *Nordic Journal of African Studies*, 17(2):23.

Benson Kituku, Wanjiku Nganga, and Lawrence Muchemi. 2021. Leveraging on cross linguistic similarities to reduce grammar development effort for the under-resourced languages: a case of Kenyan Bantu languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 83–88. IEEE.

Prasanth Kolachina and Aarne Ranta. 2019. Bootstrapping ud treebanks for delexicalized parsing. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 15–24.

Laurette Marais and Laurette Pretorius. 2023a. Extending the usage of adjectives in the Zulu AfWN. In *Proceedings of the 12th Global Wordnet Conference*, pages 303–314, Donostia, Spain.

Laurette Marais and Laurette Pretorius. 2023b. Parsing IsiZulu Text Using Grammatical Framework. In *Distributed Computing and Artificial Intelligence, Special Sessions I, 20th International Conference*, pages 167–177, Cham. Springer Nature Switzerland.

Sthembiso Mkhwanazi and Laurette Marais. 2024. Generation of segmented isiZulu text. *Journal of the Digital Humanities Association of Southern Africa*, 5(1).

Carmen Moors, Ilana Wilken, Karen Calteaux, and Tebogo Gumede. 2018. Human language technology audit 2018: analysing the development trends in resource availability in all South African languages. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*, SAICSIT '18, page 296–304.

Owen G. Mordaunt, Paul A. Williams, and Z.T. Motsa Madikane. 2023. What sets Siswati apart from isiZulu? *American International Journal of Humanities and Social Science*, 8(1):47–55.

Peter Nabende, David Bamutura, and Peter Ljunglöf. 2020. Towards Computational Resource Grammars for Runyankore and Rukiga. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*.

Finex Ndhlovu. 2022. Pan-African identities and literacies: The orthographic harmonisation debate revisited. *South African Journal of African Languages*, 42(2):207–215.

Aarne Ranta. 2009. The GF resource grammar library. *Linguistic Issues in Language Technology*, 2.

Aarne Ranta. 2011. *Grammatical framework: Programming with multilingual grammars*, volume 173. CSLI Publications.

Aarne Ranta, Krasimir Angelov, Normunds Gruzitis, and Prasanth Kolachina. 2020. Abstract Syntax as Interlingua: Scaling Up the Grammatical Framework from Controlled Languages to Robust Pipelines. *Computational Linguistics*, 46(2):425–486.

P.C. Taljaard and S.E. Bosch. 1988. *Handbook of IsiZulu*. J.L. Van Schaik.

P.C. Taljaard, J.N. Khumalo, and S.E. Bosch. 1991. *Handbook of SiSwati*. J.L. Van Schaik.

Early Child Language Resources and Corpora Developed in Nine African Languages by the SADiLaR Child Language Development Node

Michelle Jennifer White[1,2], Frenette Southwood[2], Sefela Yalala[3] and the South African Communicative Development Inventory team*
University of Plymouth[1], Stellenbosch University[2], Northwestern University[3]
michelle.white@plymouth.ac.uk, fs@sun.ac.za, sefelayalala2027@u.northwestern.edu

Abstract

Prior to the initiation of the project reported on in this paper, there were no instruments available with which to measure the language skills of young speakers of nine official African languages of South Africa. This limited the kind of research that could be conducted, and the rate at which knowledge creation on child language development could progress. Not only does this result in a dearth of knowledge needed to inform child language interventions but it also hinders the development of child language theories that would have good predictive power across languages. This paper reports on (i) the development of a questionnaire that caregivers complete about their infant's communicative gestures and vocabulary or about their toddler's vocabulary and grammar skills, in isiNdebele, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, Siswati, Tshivenda, and Xitsonga; and (ii) the 24 child language corpora thus far developed with these instruments. The potential research avenues opened by the 18 instruments and 24 corpora are discussed.

Keywords: Communicative development inventory, child language, gesture, vocabulary, grammar

1. Introduction

The dearth of instruments with which to measure early child language development in African languages and of child language corpora in these languages need attention for three main reasons. The first is that life chances are influenced by educational attainment, which requires good literacy, and that the latter is built on adequate language skills (Catts et al., 1999). It is pertinent to identify children who have poor language skills early so that they can receive the intervention necessary for the improvement of said skills (Fricke et al., 2013), and for such identification, one needs reliable measuring instruments and developmental norms. The second reason is related to the first: Child language intervention programmes need to be evidence-based and take typical child language development into account. To gain contextually relevant knowledge on typical child language development, we require instruments with which to measure and track development, and corpora to analyse so that we can answer our child language related research questions. The third reason is that most of what we know about child language development is based on research of English and other European world languages (such as German and French), and that this research (which could present a skewed picture of child language development) is what informs theories of child language development (Kidd and Garcia, 2022). To generate knowledge on child language development in African languages with which to test the generalisability of existing child language theories, we need appropriate child language measuring instruments and sizeable child language corpora in African languages.

In this paper, we report on instruments and corpora developed for isiNdebele, isiXhosa, isiZulu, Sesotho, Sesotho sa Leboa, Setswana, Siswati, Tshivenda, and Xitsonga[1] by a multilingual, multi-site team of linguists, speech-language therapists, and African language specialists. Specifically, one infant and one toddler version of a child language assessment instrument – the MacArthur-Bates Communicative Development Inventory (CDI, Fenson et al., 2007) – was developed for each of these nine official spoken African languages of

* SA-CDI team: Monicca Bhuda (University of Mpumalanga), Nina Brink (North-West University), Heather Brookes (Stellenbosch University), Nomfundo Buthelezi (University of KwaZulu-Natal), Carmen Coetsee (Stellenbosch University), William Jiyana (University of Mpumalanga), Portia Khumalo (Stellenbosch University), Babalwa Ludidi (University of Cape Town), Patricia Makaure (Stellenbosch University), Martin Mössmer (University of Michigan), Muzi Matfunjwa (North-West University), Lufuno Miriri (University of Limpopo), Mikateko Ndhambi (Sefako Makgatho Health Science University), Sibusiso Ndlangamandla (University of South Africa), Helena Oosthuizen (Stellenbosch University), Nomsa Skosana (North-West University), and Katie Alcock (Lancaster University).

[1] The corpora are stored by SADiLaR but have not yet been made available to other researchers. Enquiries about the final versions of the instruments can be directed to the second author.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 86–93
25 May, 2024.

South Africa[2], and (ii) 24 corpora were built, or are in the process of being built, with these instruments. These corpora comprise the infant and the toddler CDI data as well as transcribed language samples collected from toddlers who speak one of these nine languages. We also discuss the research possibilities these instruments and corpora afford us.

2. The South African Communicative Development Inventories

The MacArthur-Bates CDI was first developed for American English (Fenson et al., 1993) but has since been adapted for more than 100 languages from different language families (see https://mb-cdi.stanford.edu/adaptations.html), under license of, and following the guidelines of, the MacArthur-Bates Board in order to render culturally and linguistically appropriate adaptations rather than mere translations. There are two age versions of the CDI, one for infants (8 to 18 months) and another for toddlers (16 to 30 months), with the 16- to 18-month overlap being intentional[3]. CDIs are caregiver reports: The parents or other primary caregivers check off on a list which language items a child has acquired. Both the infant and toddler CDIs focus on vocabulary. Words from more than 20 semantic domains are listed alphabetically (see Table 2 further below for the domains included in the South African CDIs). Caregivers are asked to indicate which of these words the child knows. On the infant version, a list of approximately 400 words (see Table 2 for precise numbers) can be marked off for either comprehension, or comprehension *and* production. On the toddler version, a list of approximately 700 words can be marked off, for production only. The infant CDI also contains checklists for gestures, play routines, actions, and comprehension of commonly used phrases (see Section 2.2.1), whereas the toddler CDI has grammar checklists for morphology, word combinations, and sentence complexity (see Section 2.2.3).

2.1 Method for Developing the Communicative Development Inventories

2.1.1 General Protocol

Following the MacArthur-Bates Board's guidelines, research teams have utilised a range of methods to adapt CDIs to new languages (see Jarůšková et al., 2023). Many teams make use of Wordbank (Frank et al., 2017), an open access repository of CDI data, to examine which words other CDIs have included. Due to Wordbank only having come into being after the commencement of the current study, this approach was not applied. Another common way to begin the adaptation process is to translate an existing CDI into the target language (Jarůšková et al., 2023), which is subsequently expanded by adding words that are culture-specific and/or language-specific. We began by translating the American English CDI to the target languages. Following, for example, Anđelković et al. (2017) for Serbian and Jackson-Maldonado et al. (1993) for Mexican Spanish, we made use of caregiver interviews to uncover which actions, gestures, and words in the translation might be irrelevant or missing. We also employed focus group discussions and spontaneous language samples to the same effect, as discussed below.

Due to the nature of the differences between the grammars of English and African languages, the grammar sections could not use a translation of the American English CDI as their point of departure. As we will explain below, we consulted the Kiswahili and Kigiryama CDIs (Alcock et al., 2015), the limited literature available on early language development in Bantu languages, caregivers of young children speaking the relevant languages, focus groups, and our recordings of toddlers' spontaneous language samples to create a first version of the grammar section.

The main aims of this pre-pilot phase were to check for completeness and eliminate cultural bias before piloting the CDIs. Below, we discuss the steps that were followed during the adaptation process in more detail.

2.1.2 Testing the First and Second Draft Versions of the CDI

As a first step, the American English CDI was translated by three mother tongue speakers per language for isiXhosa, Sesotho, Setswana and Xitsonga. Initially, funding could only be secured for four languages, and these four were selected because we had an existing network of mother-tongue-speaking researchers available for them. The adaptation process for the remaining five languages (isiZulu, isiNdebele, Sesotho sa Leboa, Siswati and Tshivenda) was initiated two years later, after the CDIs for the first four languages had been piloted twice, and once

[2] We developed similar instruments and resources for the two official Germanic languages of South Africa, i.e., Afrikaans and South African English, but we report on those developed for the Bantu languages only and not on those developed for Afrikaans, which can also be viewed as an African language. Such resource development for South African Sign Language is yet to commence.

[3] There is also a CDI-III for children of 30 to 37 months (see https://mb-cdi.stanford.edu/cdi_iii_form.html). It is a very short questionnaire, and few research teams have developed this CDI age version for their language(s).

further funding had been secured. A main consideration was to harmonise all nine CDI language versions so that they would be comparable and allow for crosslinguistic comparisons and data pooling during future research. Considering that the first four languages' adaptations performed very well during the two pilots, the adaptation of the CDI for the last five languages did not start with a translation of the American English CDI but rather with that of a more closely related language's CDI: For the Nguni languages (isiZulu, isiNdebele and Siswati), the isiXhosa adaptation of the CDI was translated; Sesotho sa Leboa used the Sesotho and Setswana CDIs; and Tshivenda used the Xitsonga CDI. Harmonisation across languages commenced before the first pilot of the first four languages and was further refined after both the first and second pilots.

As indicated above, two rounds of piloting were completed for the first four languages, but only one for the remaining five languages because they were closely based on the four already piloted language versions of the CDI. In the first pilot, each of the preliminary adaptations of the CDIs were completed by 40 caregivers of infants 8 to 18 months old and another 40 caregivers of toddlers 16 to 32 months old. They completed paper copies of the CDIs with the help of fieldworkers who were recruited via Early Childhood Development centers and researcher networks. After the first pilot, some items were removed or replaced based on the caregiver responses.

For the second pilot of the first four languages and the only pilot of the last five languages, online CDIs were used instead of paper-based versions. The online CDIs were built on Qualtrics (Provo, Utah), eliminating possible human error in data capturing. Additionally, the Qualtrics application allowed for the collection of data without the need for internet connectivity, which is a necessity in rural areas and during the frequent electricity blackouts South Africa has been experiencing.

Data from caregivers of more than 100 infants and 100 toddlers per language was collected (see Tables 4 and 5 for exact numbers per language) during Pilot 2 of the first four languages and the only pilot of the remaining five, again with the assistance of fieldworkers, either face-to-face or (when COVID-19 social distancing regulations were in place) telephonically. This was done for respondent comfort, given that many caregivers were not able to complete the CDI themselves due to low literacy levels or technology-related limitations.

2.1.3 Development of the Actions and Gestures Section (Infant CDI)

Actions and gestures that are on the American English CDI were used as a starting point for this section. Those items which were not relevant to our context were excluded or modified after translation, and actions and gestures typically used by speakers of the target languages were added. To make the items relevant to the South African context, some had to be adapted. For example, rather than asking whether the child waved to say hello, as in the American English CDI, we asked whether the child used a gesture to greet such as waving, thumbs up, high five or something culturally similar. This was done to cover the variation that exists in children's first social gestures for greeting across the languages concerned. Imitating adult actions were also changed to be more culturally and/or contextually appropriate. For instance, brooms are used more often for cleaning than vacuum cleaners, therefore sweeping was added to the American English CDI's question about whether the child imitates adults by attempting to mop or vacuum clean.

2.1.4 Development of the Words Section

The translated CDIs were presented to individual language practitioners of each language (e.g., linguists or speech-language therapists) whereafter two focus groups per language were consulted. They consisted of professionals who work with children as well as parents of young children. The feedback from the language practitioners and focus groups led to the removal and addition of some words and/or synonyms. Words which are not relevant to everyday South African life, such as *snow suit*, were removed, whereas words had to be added when, for instance, a single word on the American English CDI could be translated in multiple ways. Consider, for example, porridge, which is a staple food for many South Africans. Various types of porridge (e.g., maize meal porridge or oatmeal porridge) can be referred to with one word, *porridge*, in English but require several words in the African languages concerned, depending on its ingredients and consistency, including *papa, mahleu*, *motoho*, or *mabele* in Sesotho, and *motogo*, *bogobe* or *phaletšhe* in Setswana. All these words for porridge were added to the word lists. When adapting the word lists, dialects or varieties of the specific languages were also considered. For this reason, the focus groups comprised of people speaking various dialects or varieties of the language in question and focus group members were requested to point out those items which were highly dialectal or variety specific.

Subsequently, 30-minute samples of naturally occurring spontaneous language were collected from six toddlers (27 to 32 months) per language. Words that were found to occur in the language samples but were not yet on the word lists were added.

Although the same protocol was followed for all languages in an attempt to facilitate crosslinguistic comparisons, the final number of words varied across languages (see Table 2). This is due, for instance, to some words being polysemous in one language while several related words were required in another.

2.1.5 Development of the Grammar Section (Toddler CDI)

Only a limited number of studies have been conducted on grammar development in children learning Bantu languages, yielding very little available empirical evidence. Such evidence on early acquired grammatical constructs was available for only a few languages (see Demuth, 2003 for a summary), namely Sesotho (e.g., Connelly, 1984; Demuth, 1992), Siswati (Kunene, 1979), isiZulu (Suzman, 1991) and Setswana (Tsonope, 1987; 1993). Thus, the grammar section had to be developed based on this limited existing literature, the language samples referred to above, and the Kiswahili and Kigiriama CDIs (Alcock et al., 2015). These were the only published full CDIs that had been adapted into African languages at the time. The grammar section of the Kiswahili/Kigiriama CDIs appeared to perform well (Alcock et al., 2015) and were thus deemed reliable for use as a starting point. Their structure was followed, yielding grammar sections which each consisted of four subsections, namely small parts of words, word complexity, word combinations, and sentence complexity (see Table 3 for more information). The language professionals and focus groups commented on the preliminary items and were encouraged to suggest examples of constructions that children acquiring the languages are likely to hear or to produce.

Across the languages, there were many similarities but also some distinct grammatical differences. The decision was made to include additional, language-specific items (more than would be needed in the final version of the CDI) for the first pilot, even if the type of construction did not occur in all the languages concerned. This was done to ascertain which items would be most effective because so little data is available on these languages. In Sesotho and Setswana, for instance, there is irregular verb inflection in the past tense, therefore items pertaining to this were included for these two languages only.

A feature common to Bantu languages is that of having several noun classes (which take the form of prefixes), with different numbers in each language. Moreover, some languages and language varieties have pre-prefixes that do not exist in others. Examples of these items that contain structures that would likely be part of a child's early grammar had to be found. The main source of these examples were the language professionals and focus groups.

After the first pilot of the first four languages, the grammar items were improved based on the caregiver responses, and the instructions were clarified to make it easier for caregivers to understand the questions about grammar. Feedback from fieldworkers was especially important to determine what might have been confusing for the caregivers.

The second pilot was conducted with caregivers of 100 toddlers and indicated that the items were suitable; the items correlated significantly with each other and with the child's age and the child's vocabulary size, the latter measured by the word section of the CDI.

The grammar sections of the second group of languages (isiNdebele, isiZulu, Sesotho sa Leboa, Siswati, and Tshivenda), were based on the first four languages', with some adaptation. Their examples came from focus group discussions with caregivers and language professionals and from natural child language recordings. Some items were substituted because they relate to aspects that are irregular in one language but not in another, for instance; or the relevant structure differed across languages. For example, Tshivenda uses a prefix to mark past tense whereas the other languages use a suffix. These grammar sections were piloted once, with 100 caregivers per language.

2.2 Content of the Final Versions of the Communicative Development Inventories

Details of the final versions of the CDIs are summarised in the tables below. Table 1 indicates the five subsections of the actions and gesture section, and the number of items in each subsection. These subsections are (i) first communicative gestures, e.g., deictic gestures such as pointing; (ii) games and routines, e.g., clapping hands, (iii) actions with objects, e.g., drinking from a cup; (iv) pretending to be a parent, which included symbolic gestures and play schemes with a 'baby', e.g., dressing or trying to dress a doll or soft toy; and (v) imitating other adult actions, e.g., writing with a pen/pencil.

Subsection	Examples of questions (English equivalents)	No. of items
First communi-cative gestures	• Requests something by extending arm and opening and closing hand or putting their hands together • Shakes head "no"	12-14
Games and routines	• Plays a hiding game (hiding their face or whole body) • Dances	4-6
Actions with objects	• Combs or brushes own hair • Throws a ball	19

Pretending to be a parent	• Covers [a doll] with a blanket • Pushes [a doll] in a stroller/pram or carries it on his/her back	13-14
Imitating other adult actions	• Cleans with a cloth • Pretends to cook	12-14

Table 1: Number and types of items in the action and gestures sections

Table 2 provides the mean number of words per semantic domain on the Infant forms and the Toddler forms of the CDI. These differ somewhat across languages, as explained in Section 2.1.5.

Subsection	Examples of words (English equivalents)	Mean no. of items	
		Infant forms	Toddler forms
Sounds	*Woof woof, uh oh / yo*	17	17
Animal words (real or toy)	*Bee, cat, donkey*	13	30
Vehicle words	*Car, taxi/combi*	10	11
Words for toys	*Ball, game*	11	14
Food and drink	*Fruit, sourmilk, sweets*	42	69
Words for clothes	*Jersey, shorts*	17	26
Words for body parts	*Arm, eye, tummy*	22	31
Words for small household items	*Bucket, matches, spoon*	36	62
Furniture words	*Bathtub, door, television*	20	27
Outside words	*Garden, mountain, stone*	11	20
Words for places to go	*Creche/school, place, yard*	7	13
Words for people	*Child, mommy, uncle*	12	21
Words for games and routines	*It's hot, high five, please*	28	34
Action words	*Bite, go, sleep*	76	138
Describing words	*Bad, clean, yucky*	16	59
Words about time	*Today, now, morning*	4	6
Words about people and things	*His/hers, me, this*	8	19
Question words	*What, why*	6	7
Words about places	*Behind, here, under*	12	19
Words about amounts	*All, more, some*	6	12
Connecting words	*And, so*	1	5
Total		375	642

Table 2: Number of items per semantic domain of the words section, average across languages

The grammar section of the CDIs is divided into four subsections. The first concerns noun and verb affixes, representing both singular and plural noun classes and past and present tense markers. These are presented in the form of *yes/no* questions. For example, caregivers are asked the equivalent of *"Has your child started adding endings to words to show that an event has already happened?",* with two or three language-appropriate examples provided.

The second subsection asks in more detail about the use of noun class prefixes and verb affixes. The first 10 noun classes are covered as singular and plural pairs, e.g., Class 3 (singular) and Class 4 (the plural of Class 3), but there are only 8 items because, for some of the languages, (i) Classes 8 and 10 have the same prefixes (with nouns in Class 10 occurring more frequently), and (ii) Class 9 has a null prefix and/or occurred less frequently in our language samples and was thus not included. The items are presented as a trio of words with increasing complexity, i.e., a noun stem with no prefix (for instance, in isiXhosa *fazi* '(married) woman'), a noun stem with a 'shadow vowel' or place holder prefix (*mfazi*), and a noun stem with a full, correct prefix (*umfazi*); see Tsonope (1993) and Demuth (1988) for a discussion of these three stages of noun class prefix acquisition. Noun stems that exemplify each item were selected based on word frequency in the language samples. This subsection also asks about the use of verb affixes. These are presented as a pair or trio of words, again in increasing complexity, i.e., a verb stem with no affix, a verb stem with the full affix, and in some cases a middle option of a partial or incorrect affix.

The third subsection asks the caregiver the equivalent of *"Has the child started to combine words to form short sentences?",* with two

language-appropriate examples provided. If the caregiver responds affirmatively, they are asked to provide three examples of the longest sentences they heard the child say that week.

The last grammar subsection asks about the length and complexity of sentences. These are given as an example with two to four options of increasing complexity. The caregiver is asked to select the form that most closely resembles what their child would say – for example, *"I want bread"* or *"I want bread and a drink".*

The number of items per subsection is presented in Table 3, along with an indication of the types of constructions that the questions ask about.

Subsection	Types of constructions	No. of items
Small parts of words	Use of prefixes and suffixes	5
Word complexity	Noun class prefixes Verb suffixes	19-21
Word combinations	Whether the child is combining words, with 3 recent examples of the longest sentences	4
Sentence complexity	*"ball table"* vs *"ball top of table"* vs "*ball is on top of the table*"	13

Table 3: Number and types of items per category of the final grammar sections

3. Child Language Corpora

We developed three types of corpora, namely one corpus for each of the nine languages containing the caregiver responses to the infant CDIs and another containing those to the toddler CDI, as well as a corpus consisting of orthographically transcribed language samples (and their audio recordings) for six of the nine languages[4]. Each of these types of corpora is briefly discussed below.

3.1 Infant CDI Corpora

The infant CDI corpora consist of the answers that the caregivers gave to the CDI items on early communicative gestures and actions and on words that the child either comprehends or comprehends and produces, as well as background information on each infant for which the CDI was completed. The background information was on the infant's birth and medical history, general health, childcare, exposure to languages, household composition, and household resources. The data of 988 infants (approximately 110 per language) are included in the form of one searchable Excel file per language. This file contains instructions for the user and separate tabs for gestures and actions and for the vocabulary items. Table 4 contains the characteristics of the completed CDIs included in the corpus of each language.

Language	Total number	Rural (%)	Female (%)
IsiNdebele	112	62 (55.4%)	55 (49.1%)
IsiXhosa	109	53 (48.6%)	53 (48.6%)
isiZulu	99	52 (52.5%)	45 (45.5%)
Sesotho	111	58 (52.3%)	59 (53.2%)
Sesotho sa Leboa	111	74 (66.7%)	60 (54.1%)
Setswana	97	46 (47.4%)	46 (47.4%)
Siswati	117	55 (47%)	60 (51.3%)
Tshivenda	126	56 (44.4%)	63 (50%)
Xitsonga	105	82 (78.1%)	43 (41%)
Total	987	538 (54.5%)	484 (49%)

Table 4: Characteristics of the infant corpora, per language

3.2 Toddler CDI Corpora

As was the case for the infant CDI corpora, the nine toddler corpora each contain background information on the toddlers. Also included are the responses of the caregivers to the CDI items on words that the child produces (and therefore, by implication, comprehends as well) and the types of grammar constructions that the child can use. Searchable Excel files for each language contain data for 1050 toddlers (approximately 116 per language) in several tabs: As for the infant corpus, one tab contains user instructions; the others contain the background information for each child, as well as the vocabulary and grammar data. The characteristics of the completed CDIs included in the toddler corpora can be seen in Table 5.

Language	Total number	Rural (%)	Female (%)
IsiNdebele	123	63 (51.2%)	61 (49.6%)
IsiXhosa	107	57 (53.3%)	55 (51.4%)
isiZulu	115	53 (46.1%)	55 (47.8%)
Sesotho	112	57 (50.9%)	64 (57.1%)

[4] Development of a language sample corpus for each of the remaining three language (Sesotho sa Leboa, Setswana and Xitsonga) is underway and will also consist of transcribed language samples and their audio recordings.

Sesotho sa Leboa	123	72 (58.5%)	65 (52.8%)
Setswana	119	61 (51.3%)	58 (48.7%)
Siswati	128	61 (47.7%)	64 (50%)
Tshivenda	128	58 (45.3%)	62 (48.4%)
Xitsonga	95	41 (43.2%)	50 (52.6%)
Total	1050	523 (49.8%)	534 (50.9%)

Table 5: Characteristics of the toddler corpora, per language

3.3 Toddler Language Samples

For isiNdebele, isiXhosa, isiZulu, Sesotho, Siswati, and Tshivenda, there are 20 transcribed language samples. These samples were collected from 10 male and 10 female toddlers of 28 to 30 months. The language samples are video recordings of natural interaction between the toddler and (a) familiar adult(s) and/or child(ren). For each toddler, there was collectively 30 to 60 minutes of recordings (see Table 6) which in total contained at least 50 different utterances per child. Some recordings were made by the fieldworker and others by the parents, other caregivers or other adults or children. The toddlers were filmed in and/or around their homes and/or daycares during everyday activities such as indoor/outdoor play or having a meal. Most children had more than one recording, because we wanted to capture conversations in various settings, with recordings ranging in length from 1 to 60 minutes. Recordings were transcribed in CHAT format (see CHILDES; MacWhinney, 2000) in order to render them ready for analysis in CLAN (MacWhinney, 2000). Table 6 indicates the characteristics of the participants who contributed to the language sample corpora and the length of the recordings, for each of the six languages for which this corpus construction has been completed.

Language	Rural / urban	Recording length (in minutes)		
		Range per child	Mean per child	Combined per language
isiNdebele	Semi-urban	50-89	64.6	1292
isiXhosa	Urban	32-69	64.5	1281
isiZulu	Semi-urban	29-69	48.5	970
Sesotho	Rural	50-67	58.7	1174
Siswati	Semi-urban	29-69	51.9	1038
Tshivenda	Rural	60-77	62.8	1255

Table 6: Characteristics of the toddler language sample corpora, per language

4. Possible Research Uses

Although research has been conducted on child language for decades already, there are only a few well-researched languages in terms of child language development, and none of these are African languages (see Kidd and Garcia, 2022). The corpora enable one to answer a range of questions on the nature and size of the vocabulary of young speakers of African languages and on how this changes as the child ages; on the types of morphology that develops first and on how morphological development progresses between the ages of 16 and 30 months; on the mean length of the utterances of toddlers of various languages; on the relationship between child characteristics, household characteristics and child experiences on the one hand and language measures (communicative gestures, vocabulary, and grammar) on the other – for any one of the nine languages or crosslinguistically.

CDIs are used the world over as data collection instruments for research and diagnostic purposes. They allow one to measure and track language development in and of itself, but also as part of studies not pertaining to language development per se, such as studies on the effect of dialogic reading, medical treatment, or creche attendance on a child's development, of which language development forms an important part. Adding nine more language-versions of the CDI to the collection of existing CDIs significantly increases the scope of such research, allowing for the inclusion of child speakers of a wider range of languages. This enables contextually relevant research findings to be generated, which can inform contextually relevant early childhood intervention programmes.

5. Bibliographical References

Alcock, K.J., Rimba, K., Holding, P., Kitsao-Wekulo, P., Abubakar, A., and Newton, C.R.J.C. (2015). Developmental inventories using illiterate parents as informants: Communicative Development Inventory (CDI) adaptation for two Kenyan languages. *Journal of Child Language*, 42:763–785.

Anđelković, D., Ševa, N., Savić, M., and Tutnjević, S. (2014). Izveštaj roditelja kao izvor podataka o ranom razvoju dečijeg govora (Parents' report as a source of information on child language development). XX naučni skup Empirijska istraživanja u psihologij. http://empirijskaistrazivanja.org/wp-content/uploads/2016/06/Knjiga-Rezimea-EIP-2014.pdf

Catts, H.W., Fey, M.E., Zhang, X., and Tomblin, J.B. (1999). Language basis of reading and reading disabilities: Evidence from a longitudinal investigation. *Scientific Studies of Reading*, 3:331–361.

Connelly, M. (1984). *Basotho children's acquisition of noun morphology*. Unpublished Ph.D. dissertation, University of Essex, UK.

Demuth, K. (1992). The acquisition of Sesotho. In D. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition*, 3. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 557-638.

Demuth, K. (1988). Noun classes and agreement in Sesotho acquisition. In M. Barlow, C.A. Ferguson (Eds.), *Agreement in natural languages: Approaches, theories and descriptions*. CSLI: University of Chicago Press, pp. 305-321.

Demuth, K. (2003). The acquisition of Bantu languages. In D. Nurse, G. Philippson (Eds.), *The Bantu languages*. Surrey, England: Curzon Press, pp. 209-222.

Fenson, L., Bates, E., Dale, P.S., Goodman, J.C., Reznick, J.S., and Thal, D. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. Baltimore, MD: Paul H. Brookes.

Fenson, L., Marchman, V.A., Thal, D.J., Dale, P.S., Reznick, J.S., and Bates, E. (2007). *MacArthur–Bates Communicative Development Inventories user's guide and technical manual.* Baltimore, MD: Paul H. Brookes.

Frank, M.C., Braginsky, M., Yurovsky, D., and Marchman, V.A. (2017). WordBank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44:677-694. https://doi.org/10.1017/S0305000916000209

Fricke, S., Bowyer-Crane, C., Haley, A.J., Hulme, C., and Snowling, M.J. (2013), Efficacy of language intervention in the early years. *Journal of Child Psychology and Psychiatry*, 54:280-290. https://doi.org/10.1111/jcpp.12010

Jarůšková, L., Smolík, F., Chládková, K., Oceláková, Z., and Paillereau, N. (2023). How to build a Communicative Development Inventory: Insights From 43 Adaptations. *Journal of Speech, Language, and Hearing Research*, 66:2095-2117. https://doi.org/10.1044/2023_JSLHR-22-00591

Jackson-Maldonado, D., Thal, D., Marchman, V., Bates, E., and Gutierrez-Clellen, V. (1993). Early lexical development in Spanish-speaking infants and toddlers. *Journal of Child Language*, 20:523-549. doi:10.1017/S0305000900008461

Kidd, E. and Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, 42:703-735. https://doi.org/10.1177/01427237211066405

Kunene, E. (1979). *The acquisition of Swati as a first language: A morphological study with special reference to noun prefixes, noun classes and some agreement markers.* Unpublished Ph.D. dissertation, University of California at Los Angeles.

Suzman, S. (1991). *The acquisition of Zulu.* Unpublished Ph.D. dissertation, Witwatersrand University, Johannesburg, South Africa

Tsonope, J. (1987). *The acquisition of Tswana noun class and agreement morphology, with special reference to demonstratives and possessives.* Unpublished Ph.D. dissertation, State University of New York, Buffalo.

Tsonope, J. (1993). Children's acquisition of Bantu noun class prefixes. *Botswana Notes & Records*, 25(1):111-117.

6. Language Resource References

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates. https://childes.talkbank.org/

7. Acknowledgments

This study was financially supported by the South African Center for Digital Language Resources. Additional funding came from The National Research Foundation of South Africa (HSD170602236563). Preliminary work for this research was supported by The British Academy Newton Fund (NG160093) and the National Research Foundation of South Africa/Swedish Foundation for International Cooperation in Research and Higher Education (NRF/STINT160918188417). Any opinion, findings, conclusions or recommendations expressed in this material are those of the authors.

Morphological Synthesizer for Ge'ez Language: Addressing Morphological Complexity and Resource Limitations

Gebrearegawi Gebremariam [†], **Hailay Teklehaymanot** [*], **Gebregewergs Mezgebe**[†]

[†]Axum University, Institute of Technology, Department of IT, Ethiopia
[*]L3S Research Center, Leibniz University Hannover,Germany
{gideygeb,gemezgebe}@aku.edu.et,teklehaymanot @l3s.de

Abstract

Ge'ez is an ancient Semitic language renowned for its unique alphabet. It serves as the script for numerous languages, including Tigrinya and Amharic, and played a pivotal role in Ethiopia's cultural and religious development during the Aksumite kingdom era. Ge'ez remains significant as a liturgical language in Ethiopia and Eritrea, with much of the national identity documentation recorded in Ge'ez. These written materials are invaluable primary sources for studying Ethiopian and Eritrean philosophy, creativity, knowledge, and civilization. Ge'ez is a complex morphological structure with rich inflectional and derivational morphology, and no usable NLP has been developed and published until now due to the scarcity of annotated linguistic data, corpora, labeled datasets, and lexicons. Therefore, we proposed a rule-based Ge'ez morphological synthesis to generate surface words from root words according to the morphological structures of the language. Consequently, we proposed an automatic morphological synthesizer for Ge'ez using TLM. We used 1,102 sample verbs, representing all verb morphological structures, to test and evaluate the system. Finally, we get a performance of 97.4%. This result outperforms the baseline model, suggesting that other scholars build a comprehensive system considering morphological variations of the language.
Keywords: Ge'ez, NLP, morphology, morphological synthesizer, rule-based

1. Introduction

Language is one of the most important aspects of our lives, as it allows us to preserve information and pass it on orally or in writing from generation to generation (Allen, 1995).

Ge'ez is an ancient Semitic language with a unique alphabet ("ኣፃ በፃ ገፃ ደ") (Adege and Mannie, 2017; Siferew, 2013). This language played a pivotal in Ethiopia and Eritrea's cultural and religious development during the Aksumite Kingdom era. Its rich literary tradition and influence in spreading Christianity across the region are notable. Although no longer spoken colloquially after the thirteenth century, Ge'ez remains significant as a liturgical language for various religious groups. Scholars and linguists are drawn to Ge'ez for its insights into the historical evolution of Semitic languages and their connections to languages such as Hebrew, Arabic, and the modern Ethiopian and Eritrean language (Dillmann and Bezold, 2003; Desta, 2010; Abate, 2014).

Besides being the liturgical language for various religious groups in Ethiopia and Eritrea, Ge'ez remains a significant writing language for religious, historical books, and literature in the history of Ethiopia(Belcher, 2012; Scelta and Quezzaire-Belle, 2001). These written resources can be primary sources for studying Ethiopian and Eritrean philosophy, creativity, knowledge, and civilization. (Abate, 2014).

Hence, preserving the Ge'ez language becomes imperative to safeguarding Ethiopia and Eritrea's cultural and historical heritage. As the language deeply intertwined with religious practices and literature, its preservation ensures the continuity of traditions and identities across generations. Besides, preserving the Ge'ez language is crucial for maintaining religious practices and literature traditions, honoring linguistic diversity and identity, contributing to the understanding of Semitic languages' evolution, and fostering cultural pride and continuity across generations in Ethiopia and Eritrea (Desta, 2010).

However, research for this language has only started recently, and no usable technology has been developed and published until now for the Ge'ez because little consideration has been given to the language, even though it is that important. Due to this, Ge'ez is still a low-resource and endangered language(Eiselen and Gaustad, 2023; Haroutunian, 2022). In documenting endangered languages or reconstructing historical languages, understanding their morphological structure is essential for accurately representing and preserving the linguistic systems (Bisang et al., 2006). For morphologically rich languages such as Ge'ez, it is essential to develop a system that can generate all surface word forms from root words because this can serve as an input for many other NLP systems, including IR systems, spelling and grammar checking, text prediction, dictionary development,

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 94–106

POS tagging, machine translation, conversational AI, and other AI-based systems. But, it is difficult to develop AI-based systems especially for low-resourced languages such as Ge'ez, etc (Eiselen and Gaustad, 2023; Haroutunian, 2022; Gasser, 2012; Saranya, 2008; Scelta and Quezzaire-Belle, 2001; Sunil et al., 2012; Wintner, 2014).

For example, consider the search results in Table 1 to evaluate the limitation of the IR system in Ge'ez word variation.

Queries	Verb Form	Results
ፈጠነ/reTene/	Perfective	9
ይፈጥን/yreTn/	Indicative	0
ይፍጥን/yrTn/	Subjective	0
ፍጠን/rTin/	Noun	1,480

Table 1: Ge'ez queries and their results from the Google search engine

As shown in Table 1, the results obtained in each query are different, even though the queries are related and generated from the verb '**ፈጠነ**/reTene/'. In this case, the query should be given in all variants of the word forms; if not, the system will fail to retrieve the related information. However, it is inconvenient to search for all variant words (Hailay, 2013). To improve the efficiency of IR systems, it is important to create a strong relationship between the stems and their variant word forms. Thus, it is important to develop a morphological synthesizer of Ge'ez and integrate it with the IR systems to get an effective IR system.

Therefore, we proposed a rule-based Ge'ez morphological synthesizer that can play a crucial role in generating surface words from the root words according to the morphological structures of the language. This study is the first attempt to develop morphological synthesizers for the Ge'ez language, although morphological synthesizers for other languages have been developed and are available for wider usage, as stated below in the related works section. As a result, our work has made the following fundamental contributions to the scientific community:

i. We designed an algorithm based on the language's morphological rules to illustrate generating TAM and PNG features. We tried to create surface words from the lexicons. The generator uses Ge'ez Unicode alphabets without transliterating to Latin alphabets. This makes it easy to use, especially for Ge'ez learners and researchers.

ii. We prepared the first publicly available datasets for Ge'ez morphological synthesizers. Another researcher can use it.

iii. Our system gives Amharic and English meanings for the perfect verb form. Therefore, this can initiate the development of the following higher Ge'ez-Amharic, Ge'ez-Tigrinya or Ge'ez-other languages dictionary projects.

2. Related Works

One of the most popular research areas in NLP is the study of morphological synthesizers. Several research projects have been conducted in this area for various international languages using different approaches(Abeshu, 2013; Koskenniemi, 1983). Let us look at some related works.

ENGLEX was developed to generate and recognize English words using TLM in PC-KIMMO. It has three essential components, including a set of phonological (or orthographic) rules, lexicons (stems and affixes), and grammar components of the word. The generator accepts lexical forms such as **spy** + **s** as input and returns the surface word spies. The online source code is available here[1].

Jabalín was developed for both analyzing and generating Arabic verb forms using Python. They created a lexicon of 15,453 entries. This was designed using a rule-based approach called root-pattern morphology. The morphological generator accepts verb lemmas to produce inflected word forms and achieved an accuracy of 99.52% for correct words (González Martínez et al., 2013).

Using a paradigm-based approach, the Morphological Analyzer and Synthesizer for Malayalam Verbs was also developed by (Saranya, 2008). This helps in creating an English-Malayalam machine translation system.

Pymorphy2 was developed for the morphological analysis and generation of Russian and Ukrainian languages (Korobov, 2015). The system used large and efficiently encoded lexicons built from Open-Corpora and LanguageTool data. A set of linguistically motivated rules was developed to enable morphological analysis and the generation of out-of-vocabulary words observed in real-world documents.

TelMore was developed by (Ganapathiraju and Levin, 2006) to handle the morphological generation of nouns and verbs in Telugu. The prototype was designed based on finite-state automata. TelMore accepts the infinitive form for the verb types and generates the present, past, and future tenses, affirmative, negative, imperative, and prohibitive forms for all genders and numbers. In addition, (Dokkara et al., 2017) also developed a morphological generator for this language. Its computational model was developed based on finite-state techniques. The system was evaluated for a total

[1] http://downloads.sil.org/legacy/pc-kimmo/engl20b5.zip

of 503 verbs. Of these verbs, 418 words were correct, and 85 words were incorrect.

(Goyal and Lehal, 2008) developed the morphological analyzer and generator for Hindi using the paradigm approach. This system has been developed as part of the machine translation system from Hindi to Punjabi. (Gasser, 2012) developed a system that generates words for Amharic, Oromo, and Tigrinya words from the given root and affixes. This has been developed based on the concept of finite-state technology. The system produced 96% accurate results (Gasser, 2012).

A morphological synthesizer for Amharic was developed by (Lisanu, 2002) using combinations of rule-based and artificial neural network approaches. However, his study was limited to Amharic perfect verb forms. Some of the generated word forms could be more meaningful. Also, this model used a transliteration of the Amharic script into Latin before any synthesis was done. The system does not allow generation for other roots that are not registered in its database. On the other hand, words are generated as output by giving the root and suffix as inputs. This may limit the number of words the model can produce compared to the words developed by the language experts. (Lisanu, 2002).

(Abeshu, 2013) developed an automatic morphological synthesizer for Afan Oromoo using a combination of CV-based and TLM-based approaches and achieved a performance of 96.28 % for verbs and 97.46% for nouns. The study indicated that developing a full-fledged automatic synthesizer for Afan Oromoo using rule-based approaches can yield an outstanding result. And it is easy to extend the system to other parts of speech with minimal effort.

The morphological synthesizers reviewed overhead are specific to their corresponding language and cannot handle Ge'ez's morphological characteristics because Ge'ez differs from these languages. To our knowledge, no research has been conducted to develop an automatic morphological generator for the Ge'ez language. Thus, we planned to create a morphological synthesizer model that can generate the derivational and inflectional morphology of Ge'ez language verbs.

3. Ge'ez Morphology

Ge'ez language has a complex morphological structure because a single word can appear in many different forms and convey different meanings by adding affixes or changing the phonological patterns of the word (Adege and Mannie, 2017). In particular, verbs have a more complex structure than other POSs in Ge'ez. Thus, Ge'ez verbs are categorized into six principal classes in their forms labeled as perfective, indicative, infinitive, subjunctive, jussive, and gerundive verb forms. Each verb form has five stem classes, and each verb stem will inflect by adding affixes to create different word forms (Desta, 2010). Generally, there are three phases to creating variant word forms in Ge'ez, as defined in (Dillmann and Bezold, 2003). These are given below, as depicted in Figure 1:

Phase I: Stem formation

Phase II:TAM formation

Phase III: PNG formation

In Phase I, the declaration of word forms using the Tense-Mood as rows and the five stems as columns is done.

In Phase II, each surface verb form obtained from Phase I is further declared using the ten subjective pronouns by appending the subject marker suffix.

In Phase III, declarations of the word forms using the ten Object Maker Suffixes for each of the words obtained in Phase II will occur.

So, two rules for suffixing verbs govern the concatenation process of morphemes to produce the surface verb forms:

- Stem + subject-marker suffix = surface word (only with SMS)
- Stem + subject-marker suffix + object-indicator suffix = surface word (with both SMS and OMS)

Hence, we can have two verb forms, one with the only direct subject marker and the other with both subject marker and object marker suffixes, as indicated below:

ቀተለ (stem) + **ክሙ** (subject marker suffix) = **ቀተ-ልክሙ** - you killed. (Surface Form).

ቀተለ (stem) + **ክሙ** (object marker suffix) = **ቀተ-ለክሙ** - he killed you (Surface Form).

ቀተለ (stem) + **ክሙ** (SMS) + **ኒ** (OMS) = **ቀተልክሙኒ** - you killed me (Surface Form).

In this case, the subject marker suffix /-**ክሙ**/ points out that the subject is "you (2 ppm)," whereas the object marker /-**ኒ**/ indicates the object "me.". Hence, the verb /**ቀተልክሙኒ**/ indicates both the subject and the object of the verb. Hence, a single verb can be a sentence in Ge'ez because it has both subject and object indicator suffixes.

4. Methodology of the study

We have reviewed several books, research reports, journals, articles, and user manuals to grasp the morphological structure of Ge'ez verbs and to know the different techniques for designing morphological synthesizers. In addition, continuous discussions were conducted with Ge'ez experts to better understand the morphological structure of the language better and to get valuable ideas for the study.

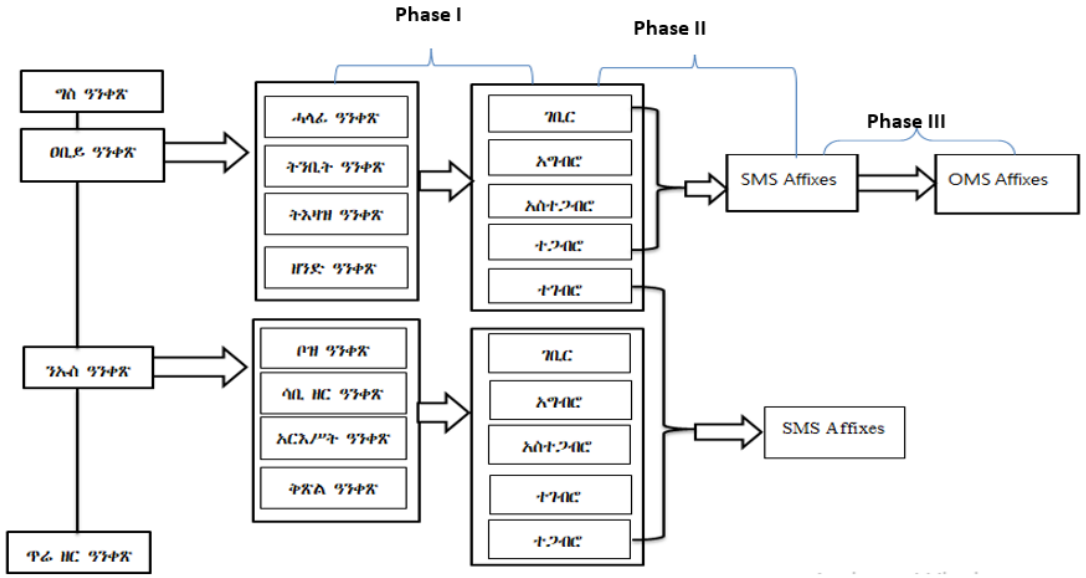


Figure 1: Phases of Ge'ez morphological word formation

4.1. Data Collection

Manually annotated data in lexicons helps test the morphological synthesizer. Since machine-readable dictionaries and word lists or an online corpus for Ge'ez were not available, the work of compiling the lexicons was started from scratch. Hence, we have compiled sample representative verbs that characterize all variations of verbs for testing and evaluating the systems's performance by consulting experts of the language. These verbs are collected from different books, like the Holy Bible, **መጽሐፈ ግስ**/Ge'ez Grammar Book/, and from lsanate siem (**ልሳናተ ሴም**) (Zeradawit, 2017). Therefore, the language lexicon prepared for this study consists of 1102 regular and irregular verbs. The affixes that can be concatenated with the verbs are also compiled into the lexicons.

4.2. Design

As defined by (Pulman et al., 1988), it is mandatory to consider at least the following basic design requirements to develop a morphological synthesizer of a language:

1. Lexicons:

Lexicon describes the list of all lemmas and all their forms. It is the heart of any natural language processing system, even though the format differs according to their needs. Consequently, the lexicons required for our study include stems, affixes, and Ge'ez alphabets. Let us see each of these lexicons in detail. i. **Stems:** In our study, the stem inputs are infinitive verb forms like **ቀቲል**/to kill/, **ሐዊር**/to walk/, **ሰጊድ**/to Prostrate/, **ፈቂድ**/to allow/, **ሐዪው**/to salivate/, etc. From these lexical inputs, the system generates inflected words for all genders and numbers by combining them with the corresponding affixes according to the set of rules of the language. The reason why we want to use the infinitive verb form as input instead of the root word/**ጥሬ ዘር**/ is to remove the ambiguity that may be created when the prototype distinguishes the input's verb category.

ii. **Affixes:** As defined by (Abebe, 2010), the affixes carry different types of syntactic and semantic information, helping to construct various words. Affixes combine with the word stems to generate various words based on the set of rules. Here, Verbal-Stem-Marker Prefixes and Person-Marker Prefixes are combined first with the input stem to generate various word stems (Abebe, 2010). Then, SMS and OMS suffixes follow in sequence. For example, consider the formation of **ይቀትለከ** /He will kill you/ using TLM in Table 2.

As indicated in Table 2, for every stem to combine with affixes, an analyzer should investigate the type of stem and the affixes that can concatenate properly to create valid surface words. Hence, a set of rules was established to handle such requirements.

iii. Ge'ez Alphabets: As described by (Koskenniemi, 1983), both the lexical and surface-level words in the two-level model are strings extracted from the language alphabets. The lexical-level strings may contain some characters that may not occur on the surface-level strings. Accordingly, Ge'ez words are constructed by the meaningful concatenation of Ge'ez alphabets. The alphabets in the Ge'ez language include all the characters starting from **ሀ**/he/ to **ፈ**/fe/ and the four other complex-compound alphabets. All the alternations of characters in the lexical strings during surface word formations are retrieved from these alphabets. Implementing these alterations is handled based on the rules in the system prototype. The two-level rules are used here to specify the permis-

sible differences between lexical and surface word representations.

2. **Morphotactics**:

Morphotactics is the model or rule of morpheme ordering that explains which classes of morphemes can follow other courses of morphemes inside a word. Ge'ez verbs have their own rules for ordering the morphemes. The order of morphemes in the word formation of Ge'ez verbs is as follows:

[Prefix] + [Prefix Circumfixes] + [stem] + [Suffix Circumfixes] + [SMS] + [OMS]

3. **Orthographic Rules:**

Orthographic rules are the spelling rules that are used to model the changes that occur in a word when two morphemes combine. Therefore, a set of rules is essential in mapping input stems to surface word forms. These rules are designed based on the morphological nature of Ge'ez for each sequence of the word formation process. Ge'ez has its own spelling rules when morphemes are concatenated with each other. For example ቀተለ/qetele/+-ኩ/ku/: ቀተልኩ/qetelku/ (here, ለ/le/ is changed to /l/ when {ቀተለ/qetele/} is added to the SMS {ኩ/ku/}).

By taking the above design requirements into account, we designed the general flow chart of the system as shown in Figure 2: As we see in the flow chart in Figure 2, the design of morphological synthesizer has the following components:

A.Stem Classifier: identifies the verb category of the stem. The classification is undertaken based on the number of heads and troops of verbs. This component also checks whether the verb stem is regular or not. Here, if the input verb contains one of the guttural alphabets (namely ሀ/he/, ሐ/He/, ኀ/H/, አ/a/ and ዐ/A/ either at their beginning or middle positions) or semi-vowel alphabets (namely የ/ye/ and ወ/we/) at any positions of the verb, it is irregular, else it is regular verb.

B.Stems Formation: This sub-component generates the various derived stems for the lexical input.

C.Signature Builder: lists the set of suffixes valid for each generated stem because every created stem has specific corresponding affixes to the stem during valid surface word formation. To establish a valid concatenation of the stems with affixes, a pattern matching mechanism is used, which is based on the notion of matching the stems with their valid affixes. For example, the word 'ይቀትል'/yqetl/ has a valid affix 'ዎ'/wo/ to create a valid word form. But, this word cannot be combined with the affix 'ክዎ'/kwo/ because the combination of the word and the affix cannot create valid word forms.

D. Boundary Change Handler: This sub-component addresses the boundary patterns occurring during the concatenation of stems and affixes based on the rules laid down on the knowledge base. These changes may be specific to every morpheme concatenation, even if these morphemes are in the same manner. Assimilation effects are occurring mostly on the boundary of the morphemes when the suffixes ከ /ke/, ኩ /ku/, ኪ /ki/, ክን /kn/ or ክሙ /kmu/ are added to the end of a verb that ends with either of the glottal alphabets, namely ቀ /QE/, ከ /ke/, or ገ /Ge/ (Lambdin, 1978). For example, observe the concatenation of the morphemes ሐደገ with ክሙ:

ሐደገ + ክሙ –> ሐደግሙ (the character ገ in ሐደገ changes to ግ and the character ክ is omitted from the morpheme ክሙ)

E. Synthesizer: This sub-component generates all possible surface word forms by concatenating the stem with the selected list of affixes using the TLM method of word generation. For example, consider the following Ge'ez word generation by TLM from Table 2:

Lexical Level	ይ	ቀ	ት	ል	+	ክ
Surface Level	ይ	ቀ	ት	ለ	0	ክ

Table 2: Generation of surface words using TLM

The rows in Table 2 depict the two-level mappings carried out during the word formation process.

F. Surface Level: Lastly, the outputs of the synthesizer are produced.

Below is our concise algorithm for producing word forms based on input lexicons:

.....

1. Start
2. Input infinitive verb stem (verb stem)
3. Classify verb regularity using classifyVerbRegularity(verbstem)
4. If regular:
4.1 For each stem in generateStems(verb stem):
4.1.1 Select affixes with selectAffixes(stem)
4.1.2 Apply boundary changes with applyBoundaryChanges(stem)
4.1.3 Concatenate changed stems with affixes
4.1.4 Print output words
5. Else (if irregular):
5.1 For each stem in generateStems (verbstem):
5.1.1 Select affixes with selectAffixes(stem)
5.1.2 Apply boundary changes with applyBoundaryChanges(stem)
5.1.3 Concatenate changed stems with affixes
5.1.4 Print output words
6. End

5. Experimentation and Evaluation

5.1. Developmental Approach

Several approaches could have been applied to developing morphological generation systems for

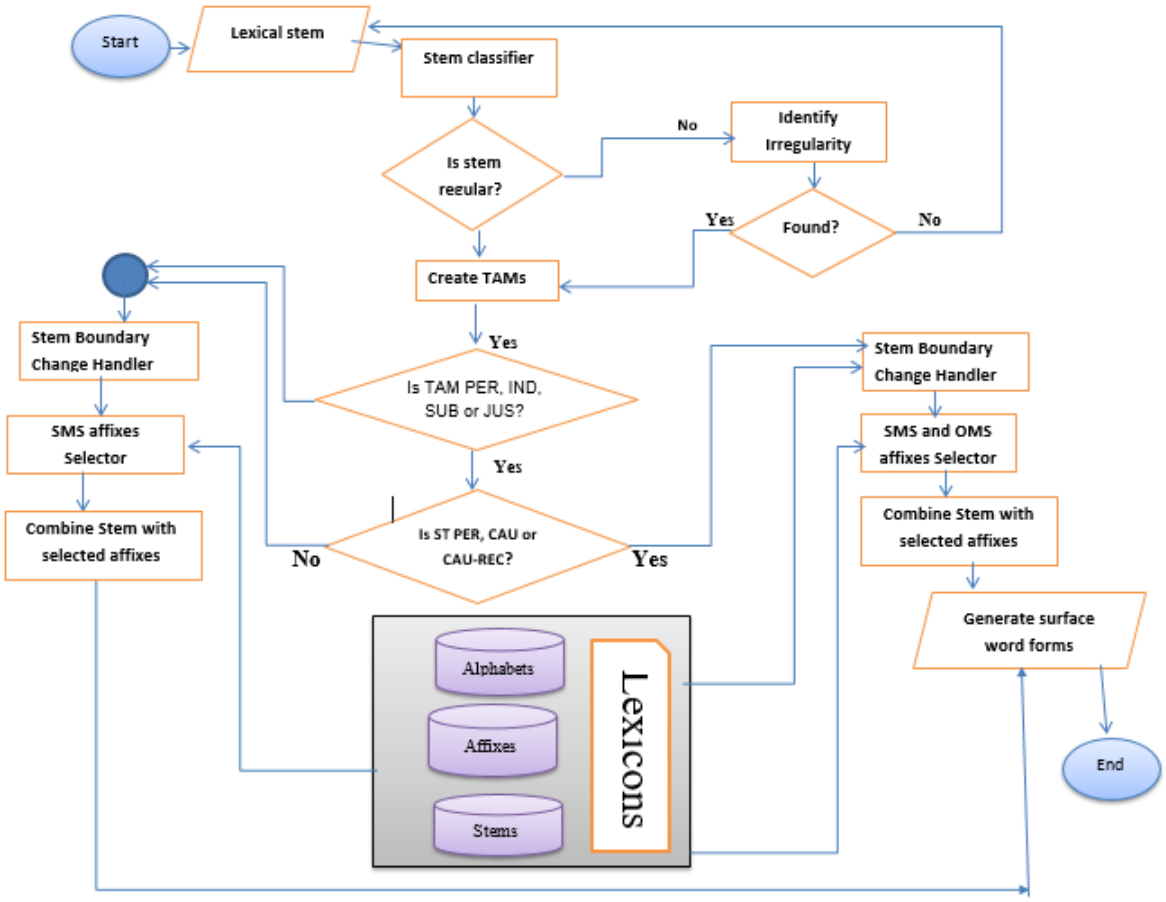


Figure 2: Flow Chart of Ge'ez morphological synthesizer

different languages. As discussed by (Kazakov and Manandhar, 2001), these approaches can be categorized as rule-based and corpus-based approaches. This study applied the rule-based approach called the Two-Level Model (TLM) of morphology to develop the prototype. TLM is used to handle the phonological and morphophonemic processes (including assimilation changes) involved in word formation (Gasser, 2011; Koskenniemi, 1984). Principally, we selected the TLM approach to map lexical entries to surface verb forms. We used the rule-based approach to develop the morphological synthesizer of the language because this approach has a faster development process with better accuracy, is more straightforward to twist, and is more accessible for formulating rules according to the language rules (Beesley and Karttunen; Shaalan et al., 2007). Moreover, the rule-based approach is practical for languages with fewer resources, such as Ge'ez, which suffers from the availability of corpora and the scarcity of data(Shaalan et al., 2010). Hence, we preferred the rule-based approach, in which a particular word is given as an input to the morphological synthesizer, and if that corresponding morpheme or root word is valid, then the system will produce surface word forms.

5.2. Testing Procedures

Systematic evaluation of the system is complex since no collected Ge'ez words are currently available for this purpose. So, to test the effectiveness of the system developed, we used the collected sample verbs. The testing procedures are as follows:

1.During the initial phase, we evaluated the system by inputting a test stem extracted from sample verbs in the lexicon, generating words, and comparing them with their expected word forms. This evaluation was conducted iteratively throughout the development of the morphological synthesizer to enhance its performance. Any errors identified during this testing, primarily related to missing rules, were rectified accordingly. Subsequent iterations of this test were conducted until satisfactory results were achieved.

2.Then, the finalized system's functionality was tested by entering sample verbs (including those with glottal or semivowel alphabets at different positions) selected by linguists.

5.3. Evaluation Procedures

Finally, we evaluated by taking regular and irregular verbs from the selected sample verbs. To evaluate the system, we used two options:

1.Manual Evaluation Using the error-counting approach, language experts manually evaluated the generated words to assess their accuracy and quality. The system accuracy is then calculated as the number of correctly generated words divided by the total number of words generated by the system multiplied by 100%.

2.Automatic Evaluation We evaluate system performance using predefined criteria and metrics without human intervention, a method akin to that described by (González Martínez et al., 2013). Subsequently, the accuracy attained from each experiment is calculated using the following formula:

$$\textbf{Accuracy} = \frac{Correctlygeneratedwords}{totalgeneratedwords} * 100 \quad (1)$$

6. Experimental Results

The accuracy assessment of the developed system involved inputting sample datasets. 7,577 words were generated from regular verbs, and 19,290 words were generated from irregular verbs. Out of these, 668 errors were identified (8 from regular verbs and 661 from irregular verbs). The accuracy rates were: 99.6% for regular verbs and 96.6% for irregular verbs. Resulting in an overall average accuracy of 97.4%. This result Surpasses the baseline(Abeshu, 2013). The percentage of words with errors was 2.6%. This promising outcome supports further research on the language. The experimental results are found and referred in the Appendix section.

7. Discussion

The system consistently produces accurate words, albeit with occasional errors. As Appendix I details, irregular verbs perform less than regular verbs, primarily due to their inherent flexibility in word-formation processes. The predominance of irregular verbs in the evaluation dataset contributes to the observed decrease in accuracy. If a more significant proportion of regular verbs were included in the evaluation, the accuracy would be expected to surpass 97.4%,given the higher accuracy rate of 99.6% observed for words generated from regular verbs.

7.1. Factors Leading to High Performance.

Despite encountering some errors, the synthesizer demonstrates remarkably high performance. This achievement can be attributed to several factors:

1.Creating correct stems Correctly generated stems generate correct surface words if the boundary changes happening during stem and affix concatenations are handled correctly. If the root stems are developed perfectly, then the words generated from these stems are correct. Hence, the performance achieved is high because most of the stems caused are right, and the boundary changes are handled correctly.

2. Handling of rules when morphemes are concatenated with each other Correct words are generated when stems and affixes are concatenated properly. For this reason, the selection of affixes for the given stem was handled properly. Therefore, handling the set of rules for word formation properly will generate valid words.

3. Handling rules for irregular word formation Ge'ez language has many irregular verbs. Irregular verbs are those that have a slight change in their morphological structure when compared to regular verbs. This is mostly happening due to the existence of one of the guttural alphabets, namely ሀ/he/, ሐ/He/, ኀ/H/, አ/a/ and ዐ/A/ either at their beginning or middle positions, or the existence of the semi-vowel alphabets, namely የ/ye/ and ወ/we/ at either position of the verbs. Irregular verbs have various rules to generate the correct word forms. These rules have slight differences from these regular word formation rules. Handling these rules of word formation gives you better accuracy. Accordingly, we have tried to handle the word formation rules as much as possible.

7.2. Error Analysis

Certain words are generated incorrectly. These errors can be attributed to the following factors:

1. Errors caused due to exceptional characters existing in the verb Some verbs have special characteristics, even though these verbs seem to have the same form as the head verb. For example, the system was designed to handle the verbs that end with the characters ቀ/qe/, ቈ/ke/, and ገ/ge/ because it is assumed that these verbs have the same morphological characteristics as other verbs. However, this may not always be true if we consider the morphological structure of the verbs ሠረቀ/šereqe/, ሐደገ/Hedege/, and ለሐቀ/leHeqe/. These verbs have differences due to the existence of guttural or semi-vowel characters, or both as shown in Table 3.

Verbs	Differences observed			
	Indicative	Subjective	Jussive	Infinitive
ሠረቀ/sereke/	ይሠርቅ/yserk/	ይሥርቅ/ysrk/	ይሥርቅ/ysrq/	ሠርቅ/seriq/
ሐደገ/Hedege/	የሐድግ/yeHedg/	ይሕድግ/yHdg/	ይሕድግ/yHdq/	ሐዲግ/Hedig/
ለሐቀ/leHeqe/	ይልሕቅ/ylHq/	ይልሐቅ/ylHaq/	ይልሐቅ/ylHaq/	ልሒቅ/lHiq/

Table 3: Errors caused by exceptional characters

As we see in table 3, the letters written in red color in the words make a difference in each word formation process even though these words are categorized in the same verb category.

2. Errors generated during concatenation of exceptional words with affixes

Some of the generated words seem to be correct both grammatically and semantically, but they

are not correct words. For example, when the morphemes ከረም/kerem/ and ነ/ne/ are concatenated, they produce the word ከረምነ/keremne/ which is the correct word. In the same way, when the morphemes አመን/amen/ + ነ/ne/, it gives አመንነ/ amenne/. However, አመንነ/ amenne/ is not the correct word. The correct word is አመነ/amene/. So, these words have different forms even though they belong to the same POS and number.

3.Errors caused due to morphological richness and varied nature of the language

This type of error occurs when testing with verbs that seems to have the same structure as other verbs in nature. But their actual output shows different word forms. For example, when we take the verbs ወለደ/welede/ and ወቀሰ/weqese/, we assume that these verbs have the same structure during the design of the prototype. But these verbs have differences in their actual word formation structure.

4. Errors caused by missing some rules The formation of the different word forms has a set of rules. Missing any of these rules generates invalid word forms. The incorrect words in Table 4 are generated because some rules and their correct forms are missing.

መራሒ ግስ (pronoun)	Incorrectly generated words	Correct words
ውእቱ (He)	ተከብበ/tekebbe/	ተከበ/tekebe/
ይእቲ (She)	ተከብበት/tekebbet/	ተከበት/tekebet/
ውእቶሙ (They-male)	ተከብቡ/tekebbu/	ተከቡ/tekebu/
ውእቶን(They-female)	ተከብባ/tekebba/	ተከባ/tekeba/

Table 4: Errors caused due to missing rules

8. Conclusion and Future Work

The study opted for the rule-based TLM approach for developing an automatic morphological synthesizer due to its simplicity, suitability, and effectiveness, especially for languages with limited corpora availability. A set of rules was meticulously designed based on expert knowledge of the language's morphological structure, forming the foundation for algorithm development from scratch to handle word formation processes. Despite the thoroughness of the morphological synthesis rules, some inaccuracies persisted in word generation, mainly stemming from the formation of invalid stems, notably for irregular verbs containing guttural and semi-vowel alphabets. Nevertheless, the prototype synthesizer exhibited promising performance, with an overall accuracy of 97.4%, indicating encouraging prospects for further research in Ge'ez linguistics. Feedback from linguists involved in the system evaluation underscored the importance of developing a comprehensive system version to enhance Ge'ez's usage and preservation within society. Recommendations were made for future researchers to address and rectify errors limiting the study's performance and to advance toward a fully functional system. Challenges encountered during the study included: A lack of Ge'ez linguistic experts. Absence of standardized references and dictionaries. Scarcity of compiled Ge'ez language lexicons. Furthermore, the complexity and agglutinative nature of Ge'ez morphology posed additional hurdles, contributing to its extensive vocabulary.

List of Acronyms

EOTC……Ethiopian Orthodox Tewahido Church
TLM……Two-Level Morphology
NLP……..Natural Language Processing
PNGs……...Persons, Numbers and Genders
POS……Parts Of Speech
TAM……Tense-Aspect-Mood
SMS……Subject Marker Suffixes
OMS……Object Marker Suffixes
IR……Information Retrieval
CV……Consonant-Vowel
PER……..Perfective
IND……..Indicative
SUB……Subjective
JUS……..Jussive
ST……Stem Type
CAU……..Causative
CAU-REC… ...Causative-Reciprocal
XML……Extensible Markup Language

References

Yitayal Abate. 2014. Morphological analysis of ge'ez verbs using memory based learning.

A. Abebe. 2010. Automatic morphological synthesizer for afaan oromoo. A thesis Submitted to School of Graduate Studies of addis ababa University in Partial fulfillment for degree masters of Science in Computer Science.

Abebe Abeshu. 2013. Analysis of rule based approach for afan oromo automatic morphological synthesizer. *Science, Technology and Arts Research Journal*, 2(4):94–97.

Abebe Belay Adege and Yibeltal Chanie Mannie. 2017. *Designing a Stemmer for Ge'ez Text Using Rule based Approach*. LAP LAMBERT Academic Publishing.

James Allen. 1995. *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc.

Kenneth R Beesley and Lauri Karttunen. Finite-state morphology: Xerox tools and techniques. *CSLI, Stanford*, pages 359–375.

Wendy Laura Belcher. 2012. *Abyssinia's Samuel Johnson: Ethiopian Thought in the Making of an English Author*. OUP USA.

Walter Bisang, Hans Henrich Hock, Werner Winter, Jost Gippert, Nikolaus P Himmelmann, and Ulrike Mosel. 2006. *Essentials of language documentation*. Mouton de Gruyter.

Berihu Weldegiorgis Desta. 2010. *Design and Implementation of Automatic Morphological Analyzer for Ge'ez Verbs*. Ph.D. thesis, Addis Ababa University.

August Dillmann and Carl Bezold. 2003. *Ethiopic grammar*. Wipf and Stock Publishers.

Sasi Raja Sekhar Dokkara, Suresh Varma Penumathsa, and Somayajulu G Sripada. 2017. Verb morphological generator for telugu. *Indian Journal of Science and Technology*, 10:13.

Roald Eiselen and Tanja Gaustad. 2023. Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct south african languages. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53.

Fitsum Gaim, Wonsuk Yang, and Jong C Park. 2022. Geezswitch: Language identification in typologically related low-resourced east african languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6578–6584.

Madhavi Ganapathiraju and Lori Levin. 2006. Telmore: Morphological generator for telugu nouns and verbs. In *Proceedings of the Second International Conference on Digital Libraries*.

Michael Gasser. 2011. Hornmorpho: a system for morphological processing of amharic, oromo, and tigrinya. In *Conference on Human Language Technology for Development, Alexandria, Egypt*, pages 94–99.

Michael Gasser. 2012. Hornmorpho 2.5 user's guide. *Indiana University, Indiana*.

Alicia González Martínez, Susana López Hervás, Doaa Samy, Carlos G Arques, and Antonio Moreno Sandoval. 2013. Jabalín: a comprehensive computational model of modern standard arabic verbal morphology based on traditional arabic prosody. In *Systems and Frameworks for Computational Morphology: Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings 3*, pages 35–52. Springer.

Alicia González Martínez, Susana López Hervás, Doaa Samy, Carlos G. Arques, and Antonio Moreno Sandoval. 2013. Jabalín: A comprehensive computational model of modern standard arabic verbal morphology based on traditional arabic prosody. In *Systems and Frameworks for Computational Morphology*, pages 35–52, Berlin, Heidelberg. Springer Berlin Heidelberg.

Vishal Goyal and Gurpreet Singh Lehal. 2008. Hindi morphological analyzer and generator. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 1156–1159. IEEE.

B. Hailay. 2013. Design and development of tigrigna search engine. A thesis Submitted to School of Graduate Studies of addis ababa University in Partial fulfillment for the Degree of Master of Science in Computer Science.

Levon Haroutunian. 2022. Ethical considerations for low-resourced machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 44–54.

Dimitar Kazakov and Suresh Manandhar. 2001. Unsupervised learning of word segmentation rules with genetic algorithms and inductive logic programming. *Machine Learning*, 43:121–162.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. In *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*, pages 320–332. Springer.

Kimmo Koskenniemi. 1983. *Two-level morphology: A general computational model for word-form recognition and production*. University of Helsinki. Department of General Linguistics.

Kimmo Koskenniemi. 1984. A general computational model for word-form recognition and production. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*. The Association for Computational Linguistics.

Thomas O. Lambdin. 1978. *Introduction to Classical Ethiopic (Ge'ez)*. Harvard Semitic Studies - HSS 24.

K Lisanu. 2002. *Design and development of automatic morphological synthesizer for Amharic perfective verb forms*. Ph.D. thesis, Master's thesis, school of Information Studies for Africa, Addis Ababa.

Stephen G Pulman, Graham J RUSSELL, Graeme D Ritchie, and Alan W Black. 1988. Computational morphology of english.

SK Saranya. 2008. Morphological analyzer for malayalam verbs. *Unpublished M. Tech Thesis, Amrita School of Engineering, Coimbatore*.

Gabriella F Scelta and Pilar Quezzaire-Belle. 2001. The comparative origin and usage of the ge'ez writing system of ethiopia. *Unpublished manuscript, Boston University, Boston. Retrieved July*, 25:2009.

Khaled Shaalan, Azza Abdel Monem, and Ahmed Rafea. 2007. Arabic morphological generation from interlingua: A rule-based approach. In *Intelligent Information Processing III: IFIP TC12 International Conference on Intelligent Information Processing (IIP 2006), September 20–23, Adelaide, Australia 3*, pages 441–451. Springer.

Khaled Shaalan et al. 2010. Rule-based approach in arabic natural language processing. *The International Journal on Information and Communication Technologies (IJICT)*, 3(3):11–19.

Muluken Andualem Siferew. 2013. *Comparative classification of Ge'ez verbs in the three traditional schools of the Ethiopian Orthodox Church*, volume 17 of *Semitica et Semitohamitica Berolinensia*. Shaker Verlag, Aachen.

R Sunil, Nimtha Manohar, V Jayan, and KG Sulochana. 2012. Morphological analysis and synthesis of verbs in malayalam. *ICTAM-2012*.

Shuly Wintner. 2014. Morphological processing of semitic languages. In *Natural language processing of Semitic languages*, pages 43–66. Springer.

A. Zeradawit. 2017. **ልሳናተ ሴም**, 1st edition. **ትንሳኤ ማተምያ ድርጅት**, Addis Ababa, Ethiopia.

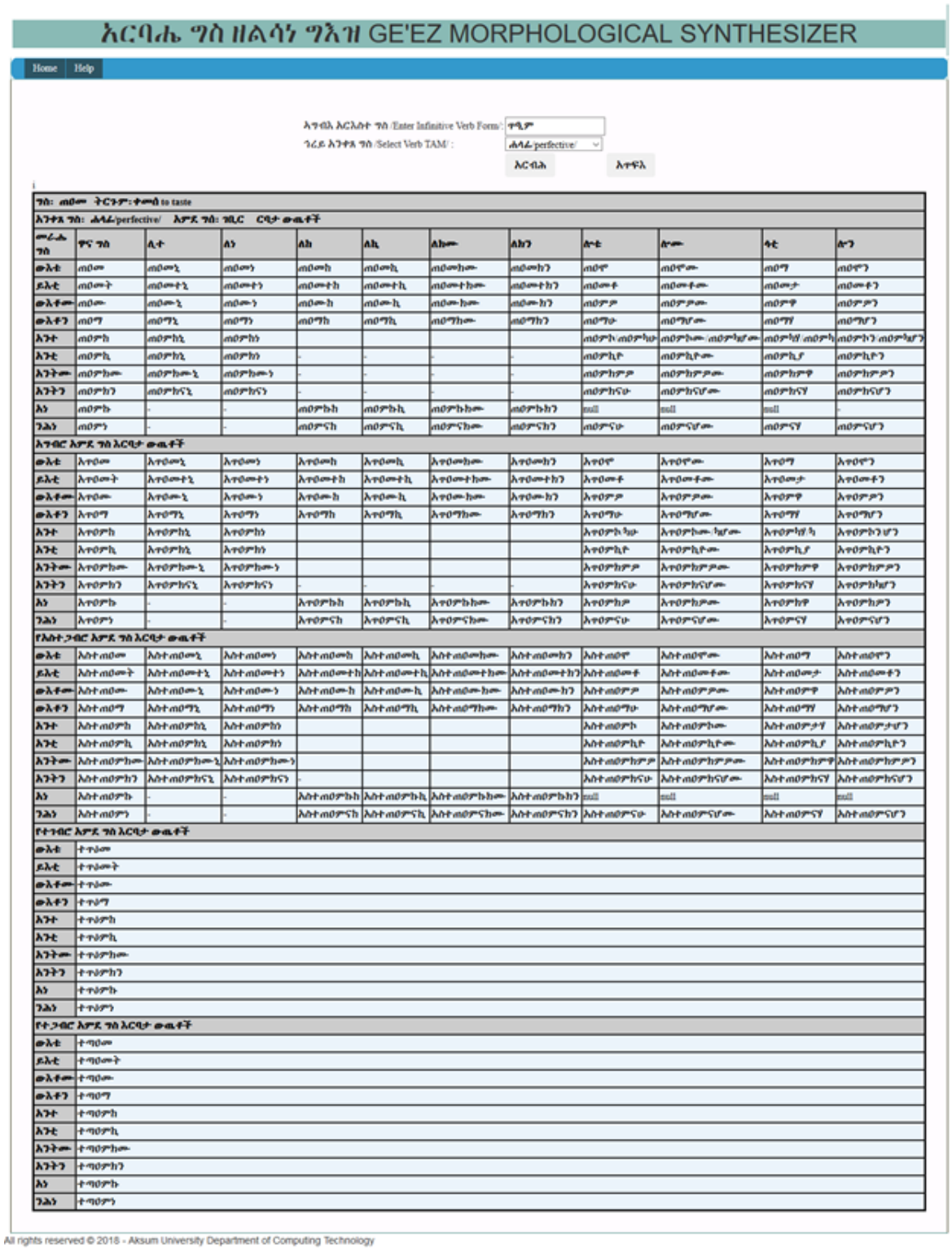
Appendix I

No.	Verb Input	Verb Form	Number of words Generated	Number of correctly Generated words	Number of wrongly Generated words	Accuracy
1.	ፈቀደ/feqede/	Regular	1269	1269	0	100%
2.	አመነ/amene/	Irregular	590	563	27	95.4%
3.	ሠረቀ/šereqe/	Irregular	1262	1262	0	100%
4.	ከደነ/kedene/	Regular	1260	1233	27	97.9%
5.	ሰበቀ/sebeke/	Irregular	1262	1262	0	100%
6.	ሐደገ/Hedege/	Irregular	1262	1162	100	92.0%
7.	መሐለ/meHele/	Irregular	580	547	33	94.3%
8.	ቀነየ/qeneye/	Irregular	580	580	0	100%
9.	አበየ/abeye/	Irregular	580	490	90	84.4%
10.	ጠወየ/Teweye/	Irregular	580	570	10	98.2%
11.	ጠዐመ/TeAme/	Irregular	1162	1162	0	100%
12.	ሐጸየ/Hetseye/	Irregular	580	490	90	84.4%
13.	ዘበጠ/zebeTe/	Regular	1262	1262	0	100%
14.	ሐመመ/Hememe/	Irregular	580	580	0	100%
15.	ወለደ/welede/	Irregular	1262	1262	0	100%
16.	ሐረደ/Herede/	Irregular	580	580	0	100%
17.	ሐለየ/Heleye/	Irregular	580	490	90	84.5%
18.	ፈደየ/fedeye/	Irregular	580	580	0	100%
19.	ከወወ/kewewe/	Irregular	580	576	4	99.3%
20.	ተለወ/telewe/	Irregular	580	580	0	100%
21.	ከበበ/kebebe/	Regular	1262	1258	4	99.7%
22.	ሐተተ/Hetete/	Irregular	580	576	4	99.3%
23.	ወጠነ/weTene/	Irregular	580	551	29	95%
24.	ረወየ/reweye/	Irregular	584	584	0	100%
25.	ለወሰ/lewese/	Irregular	1262	1262	0	100%
26.	ጸደለ/tsedele/	Regular	1262	1262	0	100%
27.	ዘረወ/zerewe/	Irregular	580	580	0	100%
28.	ገደፈ/gedefe/	Regular	1262	1262	0	100%
29.	ገረመ/gereme/	Regular	1262	1262	0	100%
30.	ወቀሰ/weqese/	Irregular	1262	1082	180	85.7%
Total			**26,867**	**26,179**	**688**	
					Average Accuracy	**97.4%**

Results obtained by the experimentation of the system prototype

Prefixes		Suffixes			Circumfixes
ኢ	ያስተ	ኩ	እ	ሆሙ	እ-እ
አ	አስተ	ነ	ኢ	ዮሙ	ን-እ
ያ	እ	ከ	ትየ	ዋ	ት-እ
ይ	ን	ኪ	ትነ	ያ	ት-ኡ
ት	እት	ክሙ	ን	ዎን	ት-ኢ
ታ	ንት	ክን	ከ	ሆን	ት-አ
ይት	ተ	አ	ሃ	ዮን	ይ-እ
ትት	ና	ኡ	ሁ	ኒ	ይ-ኡ
ታስተ	የ	አት	ዎ	ኮ	ይ-አ
ነ	ዘ	ኣ	ዮ	ቱ	
አስ	ለ	የ	ሙ	ዎሙ	
ናስተ		ኦ			

Some of the Identified Ge’ez Affixes



Screenshoot of Sample Generated words from the Synthesizer

EthioMT: Parallel Corpus for Low-resource Ethiopian Languages

Atnafu Lambebo Tonja [♠,◆,*], **Olga Kolesnikova** [♠], **Alexander Gelbukh** [♠], **Jugal Kalita** [♣],

[♠] Instituto Politécnico Nacional, Mexico, [◆] Lelapa AI,
[♣] University of Colorado Colorado Springs, USA

Abstract

Recent research in natural language processing (NLP) has achieved impressive performance in tasks such as machine translation (MT), news classification, and question-answering in high-resource languages. However, the performance of MT leaves much to be desired for low-resource languages. This is due to the smaller size of available parallel corpora in these languages, if such corpora are available at all. NLP in Ethiopian languages suffers from the same issues due to the unavailability of publicly accessible datasets for NLP tasks, including MT. To help the research community and foster research for Ethiopian languages, we introduce EthioMT – a new parallel corpus for 15 languages. We also create a new benchmark by collecting a dataset for better-researched languages in Ethiopia. We evaluate the newly collected corpus and the benchmark dataset for 23 Ethiopian languages using transformer and fine-tuning approaches.

Keywords: Parallel corpus, EthioMT, Machine Translation, low resource language, Ethiopian languages

1. Introduction

In recent years, due to advances in deep learning approaches such as the development of transformers (Vaswani et al., 2017), machine translation (MT), a core task in natural language processing (NLP), has shown dramatic improvements in terms of coverage and translation quality (Wang et al., 2021). It is well-known that a critical requirement for advancing MT is the availability of parallel corpora. The availability of parallel corpora is also necessary to facilitate the incorporation of languages in MT applications like Google Translation, Bing, and DeepL (Van der Meer, 2019). The majority of the languages in the world do not have access to such translation tools since only a few high-resource languages have received significant attention (Tonja et al., 2023b).

Most models and methods developed for high-resource languages do not work well in low-resource settings (Costa-jussà et al., 2022; Tonja et al., 2023b; King, 2015). Low-resource languages have also suffered from language technology designs (Joshi et al., 2019; Tonja et al., 2022). Creating powerful novel methods for language applications is challenging when resources are limited and only a small amount of even unlabeled data is available. The problem is exacerbated when no parallel dataset exists for specific languages (Joshi et al., 2020; Ranathunga et al., 2023; Adebara and Abdul-Mageed, 2022).

Ethiopia is a country that stands out for its remarkable cultural and linguistic diversity, with over 85 spoken languages (Woldemariam, 2007). Only a few languages of Ethiopia have received attention in the area of NLP research and application development. Most languages have been left behind due to resource limitation (Costa-jussà et al., 2022; Tonja et al., 2023b). It is hard to find publicly available datasets for Ethiopian languages to pursue NLP research because many researchers do not make their datasets publicly accessible (Tonja et al., 2023b). The unavailability of benchmark datasets and results for NLP tasks, including MT, makes research for newcomers and interested parties very difficult. This is obviously more difficult for languages with limited data in different digital forms.

This paper introduces EthioMT: a parallel corpus for low-resource Ethiopian languages paired with English, and a benchmark dataset and experimental results for 23 Ethiopian languages. Our contributions are the following: **(1)** We create a **new parallel corpus for 15 Ethiopian languages** paired with English. **(2)** We introduce the **first benchmark dataset and results for relatively better resourced Ethiopian** (Amharic, Afaan Oromo, Tigrinya and Somali) **languages**. **(3)** We evaluate MT performance with the **new corpus and present benchmark results**. **(4)** We **open-source** the parallel corpus to foster collaboration and facilitate research and development in low-resource Ethiopian languages.

2. Related work

Ethiopian languages are categorized as low-resource due to the unavailability of resources for NLP tasks, including MT (Tonja et al., 2023b). Although MT is a better-researched area for Ethiopian languages compared to other NLP applications (Tonja et al., 2023b), only a handful of languages have received adequate attention from researchers.

[*] Work done during an internship at the University of Colorado Colorado Springs.

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 107–114

Researched Languages Compared to other Ethiopian languages, the following languages have received significant attention from researchers. Nevertheless, the collected corpora are not found in one location. It is hard to find benchmark datasets in these languages and datasets and associated results to reproduce and compare MT approaches.

Amharic - Researchers have collected parallel datasets and proposed different MT approaches for Amharic-English translation (Kenny, 2018; Teshome and Besacier, 2012; Hadgu et al., 2020; Ashengo et al., 2021; Biadgligne and Smaïli, 2022; Belay et al., 2022; Gezmu et al., 2021b,a; Biadgligne and Smaïli, 2021).

Afaan Oromo - Similarly, there have been attempts to create Afaan Oromo-English MT datasets (Meshesha and Solomon, 2018; Solomon et al., 2017; Adugna and Eisele, 2010; Chala et al., 2021; Gemechu and Kanagachidambaresan, 2021).

Tigrinya - For Tigrinya-English MT, researchers have attempted to create parallel datasets (Tedla and Yamamoto, 2016, 2017; Berihu et al., 2020; Azath and Kiros, 2020; Kidane et al., 2021).

Multilingual MT Some researchers have included Ethiopian languages with other languages in multilingual MT systems. Lakew et al. (2020) collected and created benchmark results for five African languages, including those mentioned above from Ethiopia. Costa-jussà et al. (2022), Goyal et al. (2022) and Fan et al. (2021) included Ethiopian languages in their multilingual MT models and benchmark test sets. Vegi et al. (2022) crawled a multilingual parallel dataset for African languages, including Amharic and Afaan Oromo from Ethiopia.

Other languages There have been efforts to create and collect MT datasets for other Ethiopian languages. For example, Tonja et al. (2021) presented a parallel corpus for four low-resourced Ethiopian languages (Wolaita, Gamo, Gofa, and Dawuro).

3. EthioMT

3.1. Discussion of Languages

In this section, we enumerate languages included in the EthioMT corpus. Languages include in the EthioMT corpus belong to Afro-Asiatic and Nilo-Saharan language families.

3.1.1. Afro-Asiatic language family

The Afro-Asiatic language family comprises about 250 languages spoken in North Africa, parts of sub-Saharan Africa, and the Middle East. Languages belonging to this family are grouped into six sub-groups: Berber, Chadic, Cushitic, Egyptian, Omotic, and Semitic (Epstein and Kole, 1998). EthioMT contains languages belonging to the Omotic, Cushitic, and Semitic sub-groups.

1) Omotic Languages are a group of languages spoken in southwestern Ethiopia, in the Omo River region. The Ge'ez script is used to write some of the Omotic languages and the Latin script for others (Amha, 2017). Languages belonging to this group that we included in EthioMT are given below.

Basketo is spoken in the Basketo special woreda of the South Ethiopia Regional State. The Basketo language is also called Basketto, Baskatta, Mesketo, Misketto, and Basketo-Dokka. The speakers call the language "Masketo", while their neighbors call it "Basketo". The language has two dialects, Doko (Dokko) and Dollo (Dollo).

Dawuro is a language spoken by about 1.09 million people in the Dawro zone of the South West Ethiopia Peoples' Region. It is also known as Dauro, Dawragna, Dawrogna, Ometay, Cullo, or Kullo. The language has four dialects: Konta, Kucha, Longkhai, and Yawngkon.

Gamo is spoken by around 1.63 million people in the Gamo Zone of the South Ethiopia Regional State. The speakers call the language Gamotstso.

Gofa refers to the language spoken in the Gofa zone of the South Ethiopia Regional State with around 392,000 speakers.

Kafa, also known as Kefa or Kafi noono is a North Omotic language spoken in Ethiopia. It is spoken by about 830,000 people in the Keffa Zone in the South West Ethiopia Peoples' Region. The language is mainly spoken in and around the town of Bonga.

Male is spoken in the Omo Region of Ethiopia. The Male people maintain their language vigorously despite exposure to outside pressures and languages.

Shakicho, also known as Mocha, Shakacho, or Shekka, is spoken in the Sheka Zone of southwestern Ethiopia. It is closely related to Kafa. Loan words from Majang and Amharic influence the language's vocabulary.

Wolaytta is a North Omotic language spoken by the Welayta people in the Wolayita Zone of Ethiopia. It is estimated that 2 million people speak Wolaytta.

2) Cushitic languages are spoken primarily in the Horn of Africa, including Djibouti, Eritrea, Ethiopia, Somalia, and Kenya (Comrie, 2002). The Cushitic languages use the Latin and Ge'ez script. Languages belonging to this family that are included in the EthioMT group are discussed below.

Afar is spoken by the Afar people in Ethiopia, Eritrea, and Djibouti. It is also known as Afar Af, Afaraf, and Qafar af. About 1.5 million people speak Afar, the closest relative to the Saho language.

Afaan Oromo, also known as Oromo, is spoken by about 37 million people in Ethiopia, Kenya, Somalia, and Egypt. It is the third-largest language in Africa and the largest language in the Cushitic group in terms of speakers. The Oromo people are the largest ethnic group in Ethiopia and account for more than 40 percent of the population.

Language	Family	Explored prev.	No. of Speaker	Domain	Size
Afar (aar)	Afro-Asiatic / Cushitic	×	1.5M	Religious	11K
Afaan Oromo (orm)	Afro-Asiatic / Cushitic	✓	37M	Misc	**2.9M**
Awngi (awn)	Afro-Asiatic / Cushitic	×	490K	Religious	7K
Amharic (amh)	Afro-Asiatic / Ethio-Semitic	✓	57M	Misc	**1.5M**
Basketo (bst)	Afro-Asiatic/ Omotic	×	93K	Religious	7K
Dawuro (dwr)	Afro-Asiatic/ Omotic	✓	1.5M	Religious	7K
Dashenech (dsh)	Afro-Asiatic/ Cushitic	×	99K	Religious	7K
Geez (gez)	Afro-Asiatic / Ethio-Semitic	×	UNK	Religious	7K
Gamo (gmv)	Afro-Asiatic / Omotic	✓	1.09M	Religious	7K
Gofa (gof)	Afro-Asiatic / Omotic	✓	392K	Religious	7K
Gurage (sgw)	Afro-Asiatic / Ethio-Semitic	×	5.8M	Religious	28K
Hadiya (hdy)	Afro-Asiatic / Cushitic	×	1.3M	Religious	28K
Kafa (kbr)	Afro-Asiatic / Omotic	×	830K	Religious	28K
Korate (kxc)	Afro-Asiatic / Cushitic	×	500K	Religious	7K
Majang (mpe)	Nilo-Saharan / Eastern Sudanic	×	66K	Religious	9K
Male (mdy)	Afro-Asiatic / Omotic	×	105K	Religious	7K
Murule (mur)	Nilo-Saharan / Eastern Sudanic	×	300K	Religious	9K
Nuer (nus)	Nilo-Saharan /Eastern Sudanic	×	900K	Religious	29K
Shakicho (moy)	Afro-Asiatic / Omotic	×	80K	Religious	7K
Sidama (sid)	Afro-Asiatic / Cushitic	×	4M	Religious	28K
Somali (som)	Afro-Asiatic / Cushitic	✓	22.3M	Misc	**1.2M**
Tigrinya (tir)	Afro-Asiatic / Ethio-Semitic	✓	9M	Misc	**140K**
Wolaytta (wal)	Afro-Asiatic / Omotic	✓	7M	Religious	29K

Table 1: Languages and dataset details for **EthioMT** corpus. It shows languages, language families, the number of speakers, the domain, and the size of the collected dataset. In domain column **Misc** indicates *mixed* corpus collected from religious, news, and other sources. **Bold and underlined** size indicates a dataset collected from different repositories and published works and merged into one dataset for the language to create a benchmark dataset

Awngi is a Central Cushitic language spoken by about 400,000 people in northwestern Ethiopia. It is also known as Awiya, Awi, Agaw, Agau, Agew, Agow, Awawar, and Damot. Most speakers live in the Agew Awi Zone of the Amhara Region. Awngi is an Afro-Asiatic language spoken in parts of the Metekel Zone of the Benishangul-Gumuz Region.

Dashenech is also known as Dasenech, Daasanech, or Daasanach. The Daasanach people speak it in Ethiopia, South Sudan, and Kenya. The Daasanach people primarily live in the Lower Omo Valley of southwestern Ethiopia, along the eastern shore of Lake Turkana in Kenya, and in some parts of South Sudan.

Hadiya is spoken by the Hadiya people of Ethiopia. The language is also known as Hadiyyisa, Hadiyigna, Adiya, Adea, Adiye, Hadia, Hadiya, and Hadya. It is a Highland East Cushitic language. The Hadiya people are an ancient indigenous group in the southern part of Ethiopia. There are 1.4 million speakers of the Hadiya language, with 1.25 million of them speaking it as their mother tongue.

Korate is a Lowland East Cushitic language spoken by the Konso people in southwest Ethiopia. It has approximately 500,000 native speakers. The language has five dialects: Duuro, Fasha, Karatti, Kholme, and Komso. The two main dialects are Fasha and Karatti. Konso is closely related to Dirasha (also known as Gidole). It is used as a "trade language" or lingua franca beyond the area of the Konso people. The Konso people are a Cushitic ethnic group who live in large towns in south-central Ethiopia.

Sidama, or Sidaamu Afoo, is a Cushitic language spoken by the Sidama people in southern Ethiopia. It uses the Latin alphabet. Almost nine million people speak Sidama. It is the official language of the Sidama National Regional State (SNRS) and is used as a medium of instruction in primary schools. Sidama is a branch of the Highland East Cushitic family.

Somali is the official language of Somalia, spoken by 6.5 million people. It is also spoken in Ethiopia, Djibouti, and Kenya. The total number of speakers worldwide is estimated at nearly 22 million. Its closest relative is the Oromo language, spoken in parts of Ethiopia and Kenya. Other related languages include Afar and Saho.

3) Semitic languages belong to a subfamily of the Afro-Asiatic language family, including Hebrew, Aramaic, Arabic, and Ethiopic. Most scripts used to write Semitic languages are abjad. Abjad refers to an alphabetic script that omits some or all vowels. Lan-

guages belonging to this group that we study are given below.

Amharic is spoken by the Amhara and other regions in Ethiopia. It is the second most-spoken Semitic language in the world, after Arabic. Amharic is the official language of Ethiopia and has been since the 14th century. It is also spoken in other countries, including Eritrea, Canada, the United States, and Sweden. Amharic is written using graphemes called *fidal*, which means "script", "alphabet", "letter", or "character".

Ge'ez is an ancient Semitic language that originated in Eritrea and northern Ethiopia. Ge'ez is believed to be around 5,000 years old, making it older than Hebrew and other Northern Semitic languages. Orthodox and Catholic churches in Eritrea and Ethiopia still use it as a liturgical language. Ge'ez went extinct as a natural language over 1,000 years ago. It was written in two systems: an abjad and later an abugida.

Gurage is spoken by the Gurage people in central Ethiopia. The Gurage languages are written using the Ge'ez script, which is also used for other Ethiopian languages. The Gurage languages are not always mutually intelligible.

Tigrinya is spoken by about 9 million people, primarily in Eritrea and Ethiopia. It is written in the Ge'ez script, which is also used for Amharic, but the grammar and usage of Tigrinya differs significantly from Amharic.

3.1.2. Nilo-Saharan language family

Nilo-Saharan languages are a group of languages that form one of the four language families on the African continent (Dimmendaal et al., 2019). The family covers major areas east and north of Lake Victoria in East Africa and extends westward to the Niger Valley in Mali, West Africa (Comrie, 2002). Nilo-Saharan constitutes ten distinct and separate language families, including Eastern Sudanic.

Eastern Sudanic languages are a group of ten families of languages that constitute a branch of the Nilo-Saharan language family. Eastern Sudanic languages are spoken from southern Egypt to northern Tanzania. The languages used in our study by this group are given below.

Majang is spoken by the Majangir people of Ethiopia. It is a member of the Surmic language cluster, but it is the most isolated one in the group. It is classified as part of the Eastern Sudanic branch of the Nilo-Saharan language family. The Majang people live in scattered settlements in southwestern Ethiopia. They live around the urban areas of Tepi and Mett'i, southwest of Mizan Teferi and towards Gambela.

Murle is spoken by the Murle people in South Sudan and Ethiopia. The language is also known as Ajibba, Beir, Merule, Mourle, and Murule. The Murle language is part of the Surmic language family and has three dialects: Lotilla, Boma, and Olam. The Murle people number between 300,000 and 400,000. They live in Pibor County in the southeastern Upper Nile (Jonglei)

Nuer or Thok Naath is a West Nilotic language spoken by the Nuer people of South Sudan and western Ethiopia. The language is written in a Latin-based alphabet, similar to Dinka and Atuot. Over 900,000 people speak the Nuer language in diaspora communities in East Africa, Australia, and the USA.

4. Dataset

4.1. Dataset Collection

We collected datasets for 16 languages from religious domains from a website[1]. In addition to that, for Amharic, Afaan Oromo, Somali, and Tigrinya, we collected publicly available datasets (Abate et al., 2019; Lakew et al., 2020; Vegi et al., 2022) from different domains to create one benchmark dataset per language. For Dawuro, Gamo, Gofa, and Wolaita languages, we used Tonja et al. (2021) dataset to create benchmark results for fine-tuned models. A web crawler was used for each article to extract the Bible data from websites after identifying the structure of web documents. Python libraries such as requests, regular expression (RE), and Beautiful Soup (BS) were utilized to analyze website structure and extract article content from a given URL.

4.2. Sentence Alignment

After collecting the corpus for the languages, we aligned each sentence of the Ethiopian languages to a sentence in English data to prepare the dataset for the MT experiment. We followed the same procedure as Tonja et al. (2023a) to perform sentence alignment.

4.3. Dataset Pre-processing

After aligning the texts of the Ethiopian languages with their equivalent translations in English, we pre-processed the corpus before splitting it for our experiments. The pre-processing steps included removing the numeric and special character symbols, etc. We also removed parallel sentences that contain less than five words. For the baseline experiments, we split the pre-processed corpus into training, development, and test sets in the ratio of 70:10:20, respectively. Table 1 shows detailed information on selected languages, language families, domain, and their dataset size.

[1] https://www.bible.com/

5. Baseline Models

We used the following two approaches to evaluate the newly collected corpus's usability and our new benchmark dataset of four (amh, orm, som, and tir) Ethiopian languages.

The baseline transformer is a type of neural network architecture first introduced in the paper *Attention Is All You Need* (Vaswani et al., 2017). The key innovation of the Transformer architecture is the attention mechanism, which allows the network to selectively focus on different parts of the input sequence when making predictions. This contrasts traditional recurrent neural networks (RNNs), which process input sequentially and are prone to the vanishing gradient problem.

In the transformer architecture, multiple self-attention layers and feed-forward neural networks process elements of the input sequence in parallel. Each layer can be considered a "block" that takes the previous layer's output as input and applies its transformations to it. The self-attention mechanism allows the network to weigh the importance of each element in the input sequence when making predictions. In contrast, the feed-forward networks help to capture non-linear relationships among the components.

Transformers are state-of-the-art approaches widely used in NLP tasks such as MT, text summarization, and sentiment analysis. Table 3 shows parameters set up for the transformer model.

Parameters	Values
encoder_layer	6
encoder_attention_head	4
decoder_layer	6
batch_size	512
batch_type	token
decoder_attention_head	8
hidden_size	256
embed_dim	256
dropout	0.2
beam_size	5
optimizer	adam
tokenizer_type	sentencepiece
max_input_length	150

Table 2: Parameters used for transformer training

Fine tuning is the process of using a pre-trained MT model and adapting it to a specific translation task, such as translating between a particular language pair or in a specific domain. The process of fine-tuning involves taking the pre-trained model, which has already learned representations of words and phrases from a large corpus of text, and training it on a smaller dataset of specific task examples. This involves updating the pre-trained model's parameters to better capture the patterns and structures in the target translation task.

Fine-tuning can be helpful in MT because it allows the pre-trained model to quickly adapt to a new task without having to train a new model from scratch. This is especially beneficial when working with limited data or when there is a need to quickly adapt to changing translation requirements. We used **M2M100-48** a multilingual encoder-decoder (seq-to-seq) model trained for many-to-many multilingual translation (Fan et al., 2021). We used a model with 48M parameters due to computing resource limitations. We used the following parameters to fine-tune the m2m100 model.

Parameters	Values
encoder_layer	12
encoder_attention_head	16
decoder_layer	12
batch_size	512
batch_type	token
decoder_attention_head	16
hidden_size	4096
embed_dim	1024
attention_dropout	0.1
beam_size	5

Table 3: Parameters used for m2m100-48 fine-tuning

6. Results and Discussions

We evaluated the above approaches in bidirectional translation from Ethiopian languages to English and From English to Ethiopian languages. We used Sacrebleu (Post, 2018) evaluation metrics to evaluate translation models. Tables 4 and 5 show the translation results in both directions.

6.1. Using English as a source language

Table 4 shows the translation results from English to Ethiopian languages. When comparing the results of the two approaches, we observe poor performance when using a transformer rather than fine-tuning the m2m100 model. As we can see from the result, the performance of the transformer model also varies in the ranges of 0.01 – 17.8 spBLEU from language to language with different corpus sizes. This shows that a bilingual translation model trained from scratch performs poorly for low-resource language training compared to other approaches like fine-tuning multilingual translation models. Fine-tuning the multilingual model shows better results than the model built from scratch for English to Ethiopian language translation. In the fine-tuning approach, we can also observe a clear score difference between languages with larger corpora (amh, orm, tir, som) and others (e.g awn, aar, bst, etc.). This shows that fine-tuning the multilingual model will work well for languages with the largest (e.g. orm, amh) corpus sizes than languages with

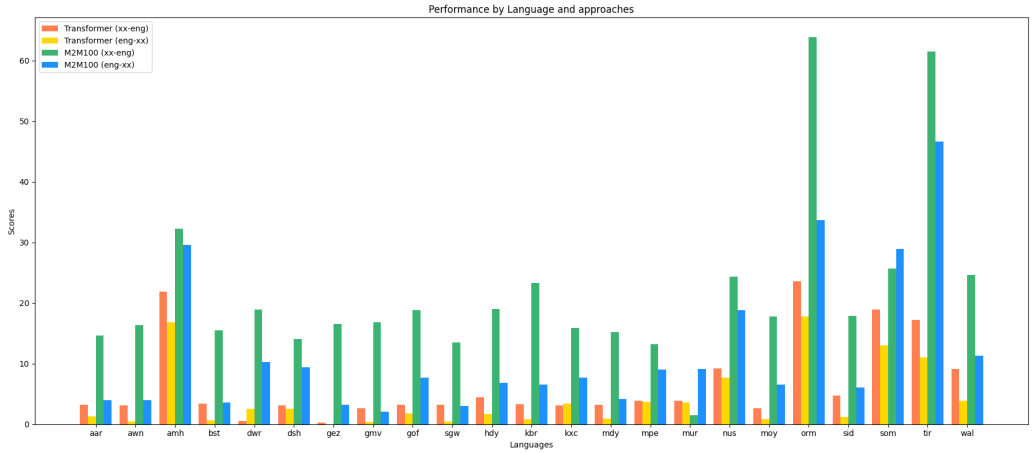


Figure 1: Benchmark translation results for transformer and fine-tuned approaches in both (from and to English/Ethiopian languages) direction

Model	en-xx																						
	aar	**awn**	**amh**	**bst**	**dwr**	**dsh**	**gez**	**gmv**	**gof**	**sgw**	**hdy**	**kbr**	**kxc**	**mdy**	**mpe**	**mur**	**nus**	**moy**	**orm**	**sid**	**som**	**tir**	**wal**	**Avg.**
	Bleu Score																							
Transformer	1.28	0.41	16.79	0.6	2.57	2.51	0.01	0.34	1.82	0.41	1.69	0.87	3.36	0.90	3.65	3.58	7.73	0.87	17.8	1.19	13.06	11.07	3.84	4.18
m2m100-fine-tuned	3.95	3.93	29.63	3.61	10.23	9.45	3.25	2.03	7.65	3.04	6.80	6.58	7.69	4.15	9.03	9.10	18.79	6.58	33.7	6.10	28.9	46.63	11.32	11.83

Table 4: Benchmark translation results from English to Ethiopian languages

Model	xx-en																						
	aar	**awn**	**amh**	**bst**	**dwr**	**dsh**	**gez**	**gmv**	**gof**	**sgw**	**hdy**	**kbr**	**kxc**	**mdy**	**mpe**	**mur**	**nus**	**moy**	**orm**	**sid**	**som**	**tir**	**wal**	**Avg.**
	Bleu Score																							
Transformer	3.18	3.14	21.9	3.39	0.52	3.07	0.28	2.68	3.21	3.18	4.42	3.26	3.14	3.21	3.91	3.92	9.23	2.63	23.6	4.77	18.9	17.2	9.16	6.60
m2m100-fine-tuned	15.61	16.32	65.34	15.47	18.92	14.11	16.57	16.79	18.79	13.52	19.04	23.27	15.90	15.20	13.26	1.48	24.40	17.78	63.9	17.86	25.71	61.50	24.62	21.79

Table 5: Benchmark translation results from Ethiopian languages to English

small (e.g. awn, bst, etc.) corpus sizes. We can also see from the results that both approaches work well for languages with mixed-domain texts compared to one domain (religion).

6.2. Using English as a target language

Table 5 shows the translation result when using English as a target language. Similarly, as we can see from the results, the transformer model performs poorly compared to the fine-tuned model when translating from Ethiopian languages to English. Compared to Table 4, translating to English shows improvements in the transformer model for similar languages. We observe that the fine-tuned model shows better Bleu scores when translating to English than when translating to Ethiopian languages. The results show that languages with large datasets have the highest performance. This shows that both models show improvements when translating from Ethiopian to English, while when translating from English to Ethiopian languages, the model is struggling with translation.

7. Conclusion and Future Works

This paper presents EthioMT, a new MT corpus for low-resource Ethiopian languages paired with English, and discusses MT experiments with results. We also present a new benchmark dataset for four Ethiopian languages collected from public repositories. We obtained benchmark results with new train, validation, and test set splits and evaluated the new corpus and new benchmark dataset using a transformer and fine-tuning multilingual translation models. From the two approaches, fine-tuning of the multilingual model outperformed the transformer approach in both translation directions.

In the future, we will work to increase the corpus sizes of the low-resource languages by extracting text from scanned documents and different sources. In addition, we will evaluate different MT approaches to low-resource languages to improve performance.

8. Bibliographical References

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021a. Extended parallel corpus for amharic-english machine translation. *arXiv preprint arXiv:2104.03543*.

Andargachew Mekonnen Gezmu, Andreas Nürnberger, and Tesfaye Bayu Bati. 2021b. Neural machine translation for amharic-english translation. In *ICAART (1)*, pages 526–532.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Asmelash Teka Hadgu, Adam Beaudoin, and Abel Aregawi. 2020. Evaluating amharic machine translation. *arXiv preprint arXiv:2003.14386*.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023a. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. *arXiv preprint arXiv:2305.17404*.

Atnafu Lambebo Tonja, Michael Melese Woldeyohannis, and Mesay Gemeda Yigezu. 2021. A parallel corpora for bi-directional neural machine translation for low resourced ethiopian languages. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 71–76. IEEE.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Mexican International Conference on Artificial Intelligence*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023b. Natural language processing in ethiopian languages: Current state, challenges, and opportunities. *arXiv preprint arXiv:2303.14406*.

Azeb Amha. 2017. The omotic language family. Cambridge University Press.

Benjamin Philip King. 2015. *Practical Natural Language Processing for Low-Resource Languages.* Ph.D. thesis.

Bernard Comrie. 2002. Languages of the world: who speaks what. In *An encyclopedia of language*, pages 529–543. Routledge.

Dorothy Kenny. 2018. Machine translation. In *The Routledge handbook of translation and philosophy*, pages 428–445. Routledge.

Ebisa A Gemechu and GR Kanagachidambaresan. 2021. Machine learning approach to english-afaan oromo text-text translation: Using attention based neural machine translation. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, pages 80–85. IEEE.

Edmund L Epstein and Robert Kole. 1998. *The language of African literature*. Africa World Press.

Gerrit J Dimmendaal, Colleen Ahland, Angelika Jakobi, and Constance Kutsch Lojenga. 2019. Linguistic features and typologies in languages commonly referred to as 'nilo-saharan'. *Cambridge Handbook of African Languages*, pages 326–381.

Haifeng Wang, Hua Wu, Zhongjun He, Liang Huang, and Kenneth Ward Church. 2021. Progress in machine translation. *Engineering*.

Hirut Woldemariam. 2007. The challenges of mother-tongue education in ethiopia: The case of north omo area. *Language Matters*, 38(2):210–235.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Lidia Kidane, Sachin Kumar, and Yulia Tsvetkov. 2021. An exploration of data augmentation techniques for improving english to tigrinya translation. *arXiv preprint arXiv:2103.16789*.

M Azath and Tsegay Kiros. 2020. Statistical machine translator for english to tigrigna translation. *Int. J. Sci. Technol. Res*, 9(1):2095–2099.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Million Meshesha and Yitayew Solomon. 2018. English-afaan oromo statistical machine translation. *International Journal of Computational Linguistic (IJCL)*, 9(1).

Mulu Gebreegziabher Teshome and Laurent Besacier. 2012. Preliminary experiments on english-amharic

statistical machine translation. In *Spoken Language Technologies for Under-Resourced Languages*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Pavanpankaj Vegi, J Sivabhavani, Biswajit Paul, Abhinav Mishra, Prashant Banjare, KR Prasanna, and Chitra Viswanathan. 2022. Webcrawl african: A multilingual parallel corpora for african languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1076–1089.

Pratik Joshi, Christain Barnes, Sebastin Santy, Simran Khanuja, Sanket Shah, Anirudh Srinivasan, Satwik Bhattamishra, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2019. Unsung challenges of building and deploying language technologies for low resource language communities. *arXiv preprint arXiv:1912.03457*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.

Sisay Adugna and Andreas Eisele. 2010. English—oromo machine translation: An experiment using a statistical approach. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Sisay Chala, Bekele Debisa, Amante Diriba, Silas Getachew, Chala Getu, and Solomon Shiferaw. 2021. Crowdsourcing parallel corpus for english-oromo neural machine translation using community engagement platform. *arXiv preprint arXiv:2102.07539*.

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Biniyam Ephrem, Tewodros Gebreselassie, et al. 2019. English-ethiopian languages statistical machine translation. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 27–30.

Surafel M Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages. *arXiv preprint arXiv:2003.14402*.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Computing Surveys*, 55(11):1–37.

Tadesse Destaw Belay, Atnafu Lambebo Tonja, Olga Kolesnikova, Seid Muhie Yimam, Abinew Ali Ayele, Silesh Bogale Haile, Grigori Sidorov, and Alexander Gelbukh. 2022. The effect of normalization for bi-directional amharic-english neural machine translation. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 84–89. IEEE.

Jaap Van der Meer. 2019. Translation technology–past, present and future. *The Bloomsbury companion to language industry studies*, pages 285–310.

Yeabsira Asefa Ashengo, Rosa Tsegaye Aga, and Surafel Lemma Abebe. 2021. Context based machine translation with recurrent neural network for english–amharic translation. *Machine Translation*, 35(1):19–36.

Yemane Tedla and Kazuhide Yamamoto. 2016. The effect of shallow segmentation on english-tigrinya statistical machine translation. In *2016 International Conference on Asian Language Processing (IALP)*, pages 79–82. IEEE.

Yemane Tedla and Kazuhide Yamamoto. 2017. Morphological segmentation for english-to-tigrinya statistical machinetranslation. *Int. J. Asian Lang. Process*, 27(2):95–110.

Yitayew Solomon, Million Meshesha, and Wendewesen Endale. 2017. Optimal alignment for bi-directional afaan oromo-english statistical machine translation. *vol*, 3:73–77.

Yohanens Biadgligne and Kamel Smaïli. 2021. Parallel corpora preparation for english-amharic machine translation. In *Advances in Computational Intelligence: 16th International Work-Conference on Artificial Neural Networks, IWANN 2021, Virtual Event, June 16–18, 2021, Proceedings, Part I 16*, pages 443–455. Springer.

Yohannes Biadgligne and Kamel Smaïli. 2022. Offline corpus augmentation for english-amharic machine translation. In *2022 5th International Conference on Information and Computer Technologies (ICICT)*, pages 128–135. IEEE.

Zemicheal Berihu, Gebremariam Mesfin Assres, Mulugeta Atsbaha, and Tor-Morten Grønli. 2020. Enhancing bi-directional english-tigrigna machine translation using hybrid approach. In *Norsk IKT-konferanse for forskning og utdanning*, 1.

Resources for Annotating Hate Speech in Social Media Platforms Used in Ethiopia: A Novel Lexicon and Labelling Scheme

Nuhu Ibrahim[†], Felicity Mulford[†], Matt Lawrence[†] and Riza Batista-Navarro[†,‡]
[†]Centre for Information Resilience, London, UK
[‡]Department of Computer Science, The University of Manchester, UK
hi@nuhuibrahim.com, {felicitym, mattl}@info-res.org, riza.batista@manchester.ac.uk

Abstract

Hate speech on social media has proliferated in Ethiopia. To support studies aimed at investigating the targets and types of hate speech circulating in the Ethiopian context, we developed a new fine-grained annotation scheme that captures three elements of hate speech: the target (i.e., any groups with protected characteristics), type (i.e., the method of abuse) and nature (i.e., the style of the language used). We also developed a new lexicon of hate speech-related keywords in the four most prominent languages found in Ethiopian social media: Amharic, Afaan Oromo, English and Tigrigna. These keywords enabled us to retrieve social media posts (also in the same four languages) from three platforms (i.e., X, Telegram and Facebook), that are likely to contain hate speech. Experts in the Ethiopian context then manually annotated a sample of those retrieved posts, obtaining fair to moderate inter-annotator agreement. The resulting annotations formed the basis of a case study of which groups tend to be targeted by particular types of hate speech or by particular styles of hate speech language.

Keywords: Hate speech, Ethiopian languages, Social media, Annotation scheme, Lexicon development

1. Introduction

Social media platforms have emerged as potent communication tools, empowering individuals to voice opinions, exchange information and participate in diverse discussions (Poell and Van Dijck, 2015). Nevertheless, the unrestricted environment of these platforms has also fostered the spread of hate speech, presenting notable hurdles to societal cohesion, particularly in culturally diverse settings like Ethiopia. With the surge of digital communication that has encouraged the intertwining of personal and public life online, hate speech has discovered novel channels for propagation, frequently targeting marginalised communities or minority groups (Kovács et al., 2021) and intensifying social divides (Targema and Lucas, 2018).

Ethiopia, a country known for its rich linguistic and cultural diversity, has witnessed the rapid spread of hate speech on social media platforms such as Twitter, Telegram and Facebook (Delelegn, 2021). Recent events, including inter-ethnic violence and political unrest, have underscored the destructive impact of online hate speech on Ethiopian society. For instance, the escalation of tensions between ethnic groups in various regions has been fuelled, in part, by the dissemination of hate speech and incendiary rhetoric on social media platforms (Delelegn, 2021).

Minority languages continue to face scarcity in computational resources for gathering and analysing extensive textual datasets, resulting in minimal to no resources for automatically detecting hate speech on social media (El-Haj et al., 2015; Kovács et al., 2021). This research aims to develop a fine-grained labelling scheme for annotating hate speech. The labelling scheme helps in producing a richly annotated hate speech dataset that does not only identify hate but also the targeted groups with protected characteristics, and the type and nature of hate speech. In addition, this research aims to develop a lexicon across four languages (Amharic, Afaan Oromo, English and Tigrigna) which are indicative of hate speech along gendered, ethnic and religious lines, which to the best of our knowledge is currently the most comprehensive one for the Ethiopian context.

This research builds upon an earlier study conducted by the Centre for Information Resilience (CIR) that considered the lived experiences and lasting impacts of online abuse through a review of existing literature and interviews with 14 women who hold prominent positions in media, civil society and other public roles in Ethiopia (Centre for Information Resilience, 2023). Their findings highlight the toxicity of online environments, and interviewees revealed that the online abuse and harassment they received have had real-world impacts, including psychological harm, damaged professional reputations, disrupted family life and the silencing of women both online and offline. Considering the gravity of hate speech proliferating on the internet in minority languages and its impact on events in Ethiopia, we argue that there is a pressing need to develop resources that will enable the development of natural language processing (NLP) methods that can aid in automatically detecting such hate speech. To this end, we present: (1) a fine-grained annotation scheme for labelling hate speech circulating in social media platforms used in Ethiopia; (2) a new

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 115–123

lexicon of hate speech-related keywords, covering inflammatory terms used in Amharic, Afaan Oromo, English and Tigrigna;[1] and (3) a corpus of social media posts annotated based on the fine-grained annotation scheme.

2. The Ethiopian Context

This research used the Ethiopian Government's definition of hate speech, as set out within the Hate Speech and Disinformation Prevention and Suppression Proclamation (No.1185/2020) (Federal Democratic Republic of Ethiopia, 2020), which defines it as "speech that deliberately promotes hatred, discrimination or attacks against a person or a discernible group of identity, based on ethnicity, religion, race, gender or disability."

In our study, we explored hate speech in social media platforms commonly used in Ethiopia. We meticulously adhered to the definition of hate speech provided by the Ethiopian Government, as stated above. Ethiopia is highly diverse in terms of languages that are in use, with over 80 languages spoken (Leyew, 2020). For reasons of feasibility and resource constraints, we decided to focus on analysing content in only four predominant languages—Amharic, Afaan Oromo, English and Tigrigna. The selection of these languages was informed by their prominence in social media platforms in the country (Zelalem, 2010).

Meanwhile, three different online platforms were chosen as the source of the content for analysis, namely X (formerly Twitter), Telegram and Facebook. These platforms were selected based on their widespread usage in Ethiopia (Daracho, 2020; Asale, 2020), coupled with the affordances provided by their policies regarding data collection and processing (Sosa and Sharoff, 2022; Giglietto et al., 2012). Additionally, these were three sites that were reported by interviewees (in CIR's earlier research) as the environments in which they faced online abuse (Centre for Information Resilience, 2023).

3. Related Work

The development of hate speech labelling schemes and lexicons for Ethiopian languages within the realm of NLP has gained increasing attention in recent years, driven by the growing recognition of the linguistic diversity and cultural richness of Ethiopia. While there is a scarcity of literature specifically dedicated to this topic, several related efforts have provided valuable insights into the challenges, methodologies and approaches relevant to hate speech labelling schemes and lexicon development for Ethiopian languages.

Peace Tech Lab (2023) reported around 21 inflammatory terms, their related spellings and associated terms, their meanings and the reasons why these terms are inflammatory. They also provided an additional 16 that are offensive and should be looked out for. Minale (2022) curated hateful keywords in Amharic and their translation in English, and then grouped the keywords into categories, namely, 'Ethiopian nation', 'gender', 'hate-related', 'offensive' and 'religious' keywords. They used these keywords to automatically collect Amharic data from three social media sites: Facebook, Twitter and YouTube. These datasets were then categorised by human annotators into four categories: 'normal speech', 'racial hate speech', 'religious hate speech', 'gender gate speech' and 'disability hate speech'.

Meanwhile, Jha and Mamidi (2017) collected sexist English posts from Twitter by matching terms or hashtags that are generally used when exhibiting what they refer to as "benevolent sexism". Some of these terms and hashtags were: *"as good as a man"*, *"like a man"*, *"for a girl"*, *"smart for a girl"*, *"love of a woman"*, *"#adaywithoutwomen"*, *"#womensday"*, *"#everydaysexism"* and *"#weareequal"*. The collected posts were manually annotated and were used to train a machine learning-based model to classify posts into three categories ('Hostile', 'Benevolent', 'Others') depending on the kind of sexism they exhibit. Similar to Minale (2022), the work by Jha and Mamidi (2017) curated Amharic sexist keywords and used them to collect posts from Twitter; they also built various classification models.

Some previous work focussed on hate speech analysis for Ethiopian languages. For instance, Getachew (2020) and Ayele et al. (2022) investigated Amharic hate speech. Kanessa and Tulu (2021) and Defersha and Tune (2021) focussed on hate speech in Afaan Oromo while Bahre (2022) studied hate speech in Tigrigna.

We found that most researchers in the Ethiopian context have concentrated on curating hate speech lexicons for Amharic, and only limited efforts have attempted to curate hate speech lexicons for other Ethiopian languages, e.g., Afaan Oromo, Tigrigna and English (as used in the Ethiopian context). In contrast, our work curated hate speech lexicons for multiple Ethiopian languages: Amharic, Afaan Oromo, English and Tigrigna. These languages are the most prominently used in social media platforms in the country (Zelalem, 2010).

Additionally, most labelling schemes developed for Ethiopian languages only classified hate speech as either 'hate' or 'no hate'. Some studies such as that by Minale (2022) went further to define cate-

[1] https://github.com/Centre-for-Information-Resilience/ethiopia-hate-speech-lexicon

gories of hate: 'normal speech', 'racial hate speech', 'religious hate speech', 'gender gate speech' and 'disability hate speech'. Jha and Mamidi (2017) also categorised hate/sexist posts into 'Hostile', 'Benevolent' or 'Others', however, they concentrated only on identifying sexism. Our research goes beyond existing work in developing a fine-grained labelling scheme that identifies three elements in hate posts: the target, type and nature of hate speech.

4. Annotation Scheme

Annotation schemes typically contain a set of guidelines or rules used to annotate or label data with specific information or attributes (Bird et al., 2009). This section discusses the development of a fine-grained labelling scheme for labelling hate speech on social media platforms in the Ethiopian context.

In line with the definition of hate speech in the Ethiopian context, stated in Section 2, for a post to be considered as containing hate speech, it has to be targeted towards an individual or group with a protected characteristic. When a post contains hate speech, our labelling scheme requires annotators to label three elements in the post: the target, type and the nature of hate speech. We refer the reader to Figure 1) for a diagram that provides an overview of the labels in our annotation scheme. In our work, the type of hate speech refers to the method of abuse (such as threats), while nature refers to the style used in the language that expresses abuse (such as irony or stereotyping).

To capture the information about the target of hate speech, the words that convey which individual/group is being targeted should be assigned any of the following labels:

- `Gender`: An individual or group of people of a particular gender.[2]
- `Ethnicity`: An individual or group of people who come from a particular place of origin and culture.
- `Religion`: An individual or group of people belonging to a particular religious group.
- `Race`: An individual or group of people possessing distinctive physical traits associated with a particular race.
- `Disability`: An individual or group of people possessing a particular disability.

[2]Although the Ethiopian Government's hate speech definition (Federal Democratic Republic of Ethiopia, 2020) does not explicitly reference sexual identity, we incorporated sexual identity within this category to capture hate based on sexual orientation.

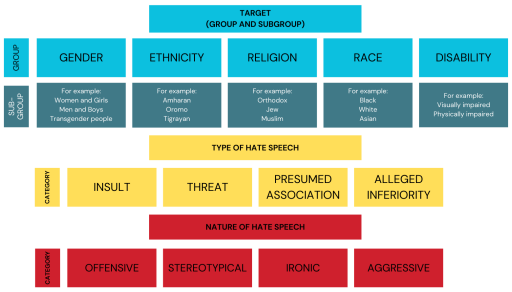


Figure 1: Overview of the categories in our annotation scheme.

Furthermore, to capture information about the type of speech, the words that convey the method of abuse should be assigned any of the following labels:

- `Insult`: Insults or denigrating expressions against an individual/group due to protected characteristics.
- `Threat`: Intimidation, threats or incitement to hatred, violence or violation of individuals' rights, due to protected characteristics.
- `Presumed Association`: Presumed association of protected characteristics with negative connotations.
- `Alleged Inferiority`: References to the alleged inferiority (or superiority) of an individual/group with a protected characteristic.

Lastly, the nature or style of hate speech often varies from one post to another. While not essential for the classification of hate speech, collecting information on style captures the nuances in the language used in expressing hate. To capture this information, hate speech-containing language needs to be labelled as any of:

- `Aggressive`: Includes strong language that seeks to physically intimidate, threaten or incite physical violence against the recipient, or which requests, suggests or promotes a violation of the recipient's rights.
- `Offensive`: Several different forms of speech, from insulting, demeaning or denigrating language, to associating the target (individual or group) with harmful or false personal traits, or suggesting the target's inferiority.
- `Ironic`: Includes jokes, satire or sarcastic messaging which targets a protected characteristic of the recipient and could be harmful. Hateful content is sometimes conveyed using

nuances in language, such as sarcasm, humour or satire.

- `Stereotypical`: Corresponds to implicit or explicit references to stereotypical beliefs or prejudices about an individual/group with protected characteristics.

The labelling scheme ensures that a multi-lingual team of annotators have a shared understanding of what labels constitute hate speech. In addition, when used, the labelling scheme produces a rich hate speech dataset that will not only tell whether a post contains hate but also the target category of the protected characteristics receiving the hate, and the type and nature of the hate received. This will help to answer substantive questions like:

- To what extent do groups with particular protected characteristics, e.g., gender, religion, ethnicity, race, etc, receive hate on social media?
- What type and nature of hate speech are prevalent?
- Do certain protected characteristics receive more hate on social media, compared to others?
- How does hate speech vary across target subgroups, i.e., women, men, homosexuals for the gender category?
- How does hate speech vary when multiple protected characteristics are targeted (i.e., hate speech that targets individuals/groups along multiple identity lines)?

5. Case Study

Our annotation scheme was applied to a case study aimed at investigating which groups with protected characteristics have often been targeted by hate speech in the Ethiopian context, as well as the type and nature of language addressed to them. This section outlines the steps we took to collect and annotate data from various social media platforms in support of the case study.

5.1. Data Collection

We collected two types of data: keywords that form a new Ethiopian hate speech lexicon, and social media posts forming a new hate speech corpus.

5.1.1. Lexicon Development

Considering the huge volume of social media posts that get published on a daily basis, we developed a lexicon of keywords to aid in the collection of posts that are likely to contain hate speech. Specifically, we collected keywords across four languages—Amharic, Afaan Oromo, English and Tigrigna—that are indicative of hate speech along gendered, ethnic and religious lines.

The lexicon was developed through desk-based research that employed both identification and refinement of existing hate speech lexicons (Minale, 2022; Degu, 2022; Getachew, 2020; James, 1998; Jha and Mamidi, 2017; Gashe, 2022; Gao et al., 2017; Peace Tech Lab, 2023; Hatebase.org, 2023; Thalikir, 2016; Centre, 2021; Shariatmadari, 2016; Center for the Advancement of Rights and Democracy, 2023), the identification of other keywords and narratives during in-person, semi-structured interviews carried out during CIR's earlier study (Centre for Information Resilience, 2023) and a roundtable discussion that brought together 21 individuals from an array of civil society organisations, UN agencies, and women and girls' rights advocacy groups.

A first draft of the lexicon was shared with partners, stakeholders and roundtable attendees in Ethiopia for feedback. It became apparent at this stage that there was confusion about why some terms had been included in the lexicon, as they may not, on their own, constitute hate speech. It was clarified to the stakeholders that the keywords will be leveraged only for collecting as many posts as possible (a high-recall but low-precision approach), and manual inspection is still necessary, as we recognise that—as with any dictionary-based approach—many keywords are ambiguous and thus their presence in a post does not necessarily mean that the post contains hate speech. Hence, human annotators will analyse whether the content indeed contains hate speech, according to the developed labelling scheme.

The resulting lexicon consists of 2,058 inflammatory keywords across the four languages within the scope of this study. We believe that, to date, this is the most comprehensive lexicon for the Ethiopian context. Figure 2 and 3 respectively show the number and distribution of keywords curated for each protected characteristic.

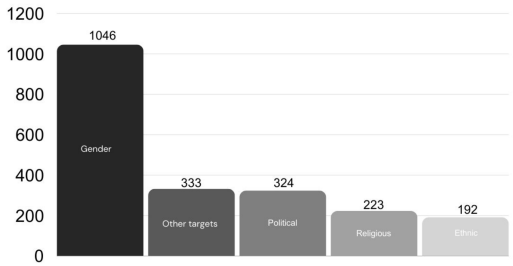


Figure 2: Number of keywords curated for each protected characteristic

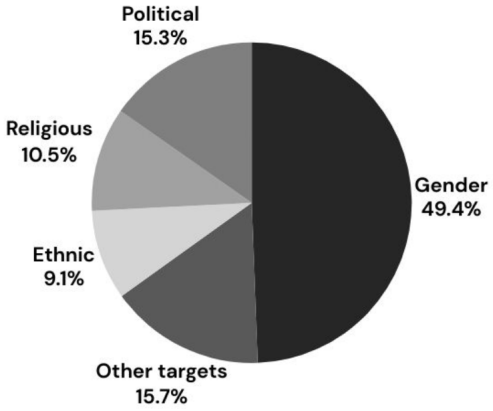


Figure 3: Distribution of keywords curated for each protected characteristic

5.1.2. Data Collection

Social media posts were collected from the platforms of interest, namely, X (formerly Twitter), Telegram and Facebook. To collect data from X, the Meltwater social media analysis tool[3] was employed. Meltwater supports the use of keyword search for tweets posted no longer than 18 months from the date of search. To ensure a relevant sample was obtained, English posts were only retrieved if they originated from Ethiopia.

In collecting data from Telegram, the official Telegram APIs[4] were used. As Telegram supports only searches within Telegram Channels to which a user belongs, social media experts from Ethiopia were engaged to meticulously curate a list of widely popular and influential public Telegram Channels in Ethiopia. Subsequently, we joined a total of 285 Telegram channels; the Telegram posts in these channels that contain keywords in our lexicon were collected.

To extract data from Facebook, social media experts from Ethiopia were again engaged in selecting a list of prominent and influential public Ethiopian Facebook groups and profiles. As an outcome of this engagement, a list of 300 Facebook profiles or groups was curated, and posts from these groups containing keywords in our lexicon were collected.

5.1.3. Data Pre-processing

Data pre-processing is a crucial step that involves cleaning, transforming and organising raw textual data to make it suitable for analysis (Tabassum and Patil, 2020).

Textual data collected from social media often contain irrelevant or erroneous information that complicates analysis or interpretation. To mitigate this issue, we carried out the following tasks on the datasets from X, Telegram and Facebook:

- Removal of HTML tags and special characters.
- Case-folding of text (i.e., making all characters lowercase) to ensure case insensitivity.
- Removal or replacement of punctuation.
- Removal of duplicate posts.

The following cleaning tasks were done only on the datasets in the English language:

- Removal of numerical values, dates and other non-textual information.
- Removal of stop words that do not carry any significant meaning (e.g., *"and"*, *"the"*, *"in"*).
- Normalisation of abbreviations and acronyms.

5.1.4. Data Anonymisation

Any usernames in the posts were anonymised in line with ethical requirements, in order to protect the privacy and confidentiality of individuals whose data is being used for research and analysis. This was done by replacing all usernames, i.e., any word appearing after the '@' symbol with the word *"USERNAME"*.

5.1.5. Sampling for Further Analysis

The data collection process resulted in the collection of tens of millions of posts, as illustrated in Table 1. Even after extensive data pre-processing, which involved removing duplicates and excessively short posts, over 5 million posts remained.

Due to the constraints posed by limited human resources available for manual annotation to determine hate content, we selected a random sample to obtain more manageable datasets. Table 1 shows the number of posts resulting from each step of the data preparation process and the number of posts chosen for subsequent analysis.

5.2. Annotation Task

The annotation task entails enlisting proficient human annotators who are familiar with the domain of interest to employ the developed labelling scheme for determining whether the posts in the collected dataset contain hate speech. The annotators used Doccano (Nakayama et al., 2018), an open-source annotation tool that we employed to label the posts

[3] https://explore.meltwater.com/meltwater-media-monitoring

[4] https://core.telegram.org/

	X	Telegram	Facebook
Posts collected	865,224	326,471,094	7,230
Posts after preprocessing	527,522	906,471	7,230
Random sample for annotation	2634	2107	2264

Table 1: The number of posts obtained in each step of the data preparation process.

in our datasets according to our annotation scheme, i.e., to annotate the hate speech targets (protected characteristics), the type and nature of hate speech.

6. Annotation Results

To ensure consistency in the application of the fine-grained labelling scheme, it was essential to calculate inter-annotator agreement (IAA) scores.

Two human annotators were enlisted to annotate the randomly chosen English posts. The primary annotator who participated in the development of the fine-grained labelling scheme and is knowledgeable of the Ethiopian context, was responsible for annotating the entire selection of English posts. To allow for estimation of IAA, a secondary annotator was assigned to annotate 10% of the dataset annotated by the primary annotator.

For Amharic, the primary annotator, a native Amharic speaker with experience in social media analysis, undertook the annotation of the entire Amharic dataset, while the other two annotators (who were assigned with the Tigrigna and Afaan Oromo datasets) were tasked with annotating approximately 10% of the dataset annotated by the primary annotator. IAA was subsequently estimated using the posts annotated by all three annotators to assess the level of agreement and consistency.

For Afaan Oromo and Tigrigna, an annotator was enlisted per language. IAA was considered unnecessary for the annotators, as these same annotators had previously worked on annotating the Amharic dataset, and the results of IAA agreement on the Amharic dataset indicated their competence in identifying hate speech, labelling its target, categorising speech types and assessing the sentiment of hate.

IAA was calculated using Cohen's Kappa (k) and Fleiss' Kappa metrics. The IAA scores, presented in Table 2, showed fair to moderate agreement between annotators (Landis and Koch, 1977).

Language	Annotators	Kappa	Agreement
English	E1 & E2	0.46	Moderate
Amharic	A1 & A2	0.38	Fair
Amharic	A1 & A3	0.46	Moderate
Amharic	A2 & A3	0.32	Fair
Amharic	A1, A2 & A3	0.39	Fair

Table 2: Result and interpretation of estimating inter-annotator agreement between annotators in terms of Kappa scores.

7. Discussion

The resulting lexicon covers a higher percentage of gender-related keywords (49.4%) compared to those related to ethnicity (9.1%) or religion (10.5%); see Figures 2 and 3. Despite this imbalance, it is worth noting that the corpus of social media posts constructed based on our lexicon nevertheless revealed a greater prevalence of hate speech targeting other identity groups, as illustrated in Figure 4. For example, out of all the posts in the full dataset, ethnic hate speech comprised 44.5%, whereas gendered hate speech and religious hate speech represented 30.2% and 17.5%, respectively. Racial hate and hate speech targeting people with disabilities made up a smaller proportion of the dataset (4.6% and 0.3%, respectively).

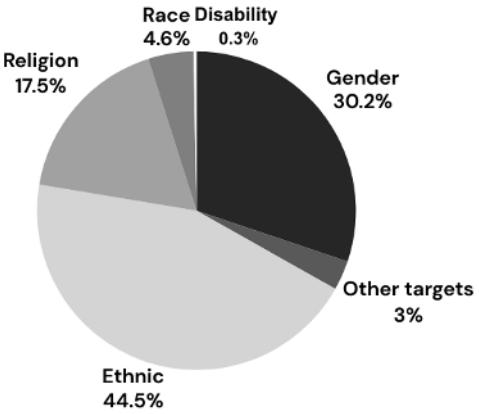


Figure 4: Distribution of hate-containing posts according to hate target.

As can be seen in Figure 5, when the protected characteristic identity groups are broken down into individual hate targets, other interesting trends become visible. The more targeted groups are women and girls (21% of the dataset), closely followed by Oromos (19.1%) and Amharans (16.7%). As the lexicon comprised more gender-related keywords, this is not surprising. Other targets of hate speech within the dataset, albeit in smaller proportions, in-

clude Orthodox Christians (8.7%), men (5.9%) and Tigrayans (5.5%).

The 'additional hate targets' category (in Figure 5) is comprised of all the other target groups outside of the top 7 most prevalent targets; this includes Protestants, white people, transgender people, atheists, Arabs, multiracial people and Jews. The 'other target' category was selected in cases where hate speech targeting a protected characteristic was present, but it does not fall under any of the categories.

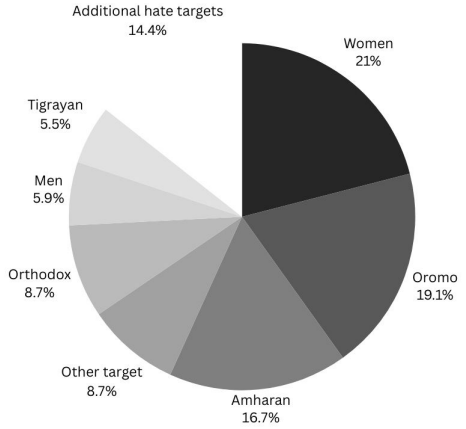


Figure 5: Distribution of hate-containing posts according to specific hate target.

Table 3 shows that women and girls receive proportionally more insulting hate speech (36.55%) than Amharans (28.31%), Muslims (29.51%) and Oromos (22.51%). They receive less insulting hate speech compared to homosexual people (38.96%) and Tigrayans (45.21%). Additionally, women and girls receive proportionally more hate containing alleged inferiority (22.2%), followed by Muslims (16.39%), Amharans (11.90%), homosexuals and Tigrayans (12.99% and 13.01, respectively), and Oromos (9.98%). Conversely, women and girls receive (proportionally) the least threats (13.51%) compared to Oromos (26.11%), Amharans (22.22%), Tigrayans (22.6%), homosexual people (20.78%) and Muslims (18.03%). Women and girls are also among the hate target subgroups which receive proportionally less hate containing presumed association (27.74%), compared to Amharans (37.57%), Muslims and Oromos (36.07% and 41.4%, respectively).

Interestingly, it was identified that offensive language is the more prevalent nature or style of hate speech across all hate targets analysed (see Table 4). Contrary to the pattern observed in offensive language, women and girls receive the highest proportion of stereotypical language (26.17%), closely followed by homosexuals (22.81%), then Muslims (12.5%), Amharans (9.27%), Oromos (4.98%)

Target	Type	%
Women	Insult	36.55
	Presumed Association	27.74
	Threat	13.51
	Alleged Inferiority	22.20
Amharan	Insult	28.31
	Presumed Association	37.57
	Threat	22.22
	Alleged Inferiority	11.90
Oromo	Insult	22.51
	Presumed Association	41.40
	Threat	26.11
	Alleged Inferiority	9.98
Muslim	Insult	29.51
	Presumed Association	36.07
	Threat	18.03
	Alleged Inferiority	16.39
Tigrayan	Insult	45.21
	Presumed Association	19.18
	Threat	22.60
	Alleged Inferiority	13.01
Homosexual	Insult	38.96
	Presumed Association	27.27
	Threat	20.78
	Alleged Inferiority	12.99
Orthodox	Insult	32.78
	Presumed Association	36.11
	Threat	24.44
	Alleged Inferiority	6.67

Table 3: Number of hate-containing posts according to target (top 7) and type of hate speech.

and Tigrayans (2.65%). Only Muslims receive a higher proportion of ironic language (21.25%) than Women and girls (20.37%). Even so, women and girls are considerably more targeted by ironic language than Tigrayans (11.5%), homosexuals (8.77%), Amharans (6.85%) and Oromos (5.32%).

8. Conclusion

In our research, we developed a fine-grained annotation scheme for labelling hate speech in posts published in social media platforms used in Ethiopia. The annotation scheme formed the basis of producing a richly annotated hate speech corpus that does not only identify hate-containing posts but also the targeted protected characteristics, the type of hate, and the nature of the language used in hate speech.

In addition, this research produced a lexicon covering four languages used in Ethiopia, i.e., Amharic, Afaan Oromo, English and Tigrigna, that contains keywords that are indicative of hate speech along gendered, ethnic and religious lines. To the best of our knowledge, this lexicon is currently the most

Target	Nature	%
Women	Aggressive	7.29
	Ironic	20.37
	Offensive	46.17
	Stereotypical	26.17
Amharan	Aggressive	23.39
	Ironic	6.85
	Offensive	60.48
	Stereotypical	9.27
Oromo	Aggressive	39.20
	Ironic	5.32
	Offensive	50.50
	Stereotypical	4.98
Muslim	Aggressive	17.50
	Ironic	21.25
	Offensive	48.75
	Stereotypical	12.50
Tigrayan	Aggressive	29.20
	Ironic	11.50
	Offensive	56.64
	Stereotypical	2.65
Homosexual	Aggressive	12.28
	Ironic	8.77
	Offensive	56.14
	Stereotypical	22.81
Orthodox	Aggressive	27.13
	Ironic	6.20
	Offensive	58.91
	Stereotypical	7.75

Table 4: Number of hate-containing posts according to target (top 7) and nature of hate speech.

comprehensive one for the Ethiopian context. Our future work will be focussed on investigating how the annotated corpus resulting from this study, can enable the development of machine learning-based models that can automatically detect and categorise hate speech, as well as automatically identify the specific targets of hate speech.

Ethics Statement

The Centre for Information Resilience (CIR) follows the Berkeley Protocol on Digital Open-Source Investigations. For this study, care was taken to anonymise data and to comply with the terms and conditions of the platforms. To mitigate the impact of vicarious trauma, annotators were offered one-to-one support from the CIR Research Coordinator (the second author of this paper). This was to ensure that the annotators were not directly impacted by exposure to hate speech. Annotators were also made aware that they have access to appropriate resources should professional help become necessary.

Acknowledgements

We would like to thank Adyam Solomon Tesfay, Alemu Teshome Baki and Fasika Tadesse for their hard work as annotators of our datasets.

This material has been funded by UK International Development from the UK government; however, the views expressed do not necessarily reflect the UK government's official policies.

Bibliographical References

Moges Ayele Asale. 2020. The Tributes and Perils of Social Media Use Practices in Ethiopian Sociopolitical Landscape. In *Proceedings of the 22nd HCI International Conference*, volume 12427, page 199, Copenhagen, Denmark. Springer Nature.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022. The 5Js in Ethiopia: Amharic Hate Speech Data Annotation using Toloka Crowdsourcing Platform. In *Proceedings of the 2022 International Conference on Information and Communication Technology for Development for Africa*, pages 114–120. IEEE.

Weldemariam Bahre. 2022. *Hate speech detection from Facebook social media posts and comments in Tigrigna language*. Ph.D. thesis, St. Mary's University.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Center for the Advancement of Rights and Democracy. 2023. CARD's Bi-weekly Social Media Conversation Sensitivity Report.

The Wilson Centre. 2021. Malign Creativity: How Gender, Sex, and Lies are Weaponized Against Women and girls Online.

Centre for Information Resilience. 2023. Silenced, shamed, and threatened: The online abuse of women who participate in Ethiopian public life.

Lisanu Damene Daracho. 2020. Social Media Impact on Social Life of Public Servant in Mari Mansa District, Dawuro Zone, Southern Region, Ethiopia. *New Media and Mass Communication*, 93:1–7.

NB Defersha and KK Tune. 2021. Detection of Hate Speech Text in Afan Oromo Social Media Using Machine Learning Approach. *Indian J Sci Technol*, 14(31):2567–78.

Mekuanent Degu. 2022. Amharic dataset for hate speech detection. Mendeley Data.

Misganaw Delelegn. 2021. *Hate Speech Regulation in Ethiopia: Lessons to Be Learned From Other Jurisdictions*. Ph.D. thesis, Bahir Dar University.

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2015. Creating language resources for under-resourced languages: methodologies, and experiments with Arabic. *Language Resources and Evaluation*, 49:549–580.

Federal Democratic Republic of Ethiopia. 2020. Hate Speech and Disinformation Prevention and Suppression Proclamation (No. 1185/2020).

Lei Gao, Alexis Kuppersmith, and Ruihong Huang. 2017. Recognizing Explicit and Implicit Hate Speech Using a Weakly Supervised Two-path Bootstrapping Approach. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 774–782, Taipei, Taiwan. Asian Federation of Natural Language Processing.

S. M. Gashe. 2022. Hate Speech Detection and Classification System in Amharic Text with Deep Learning. List of Amharic Hate Speech Keywords (Lexicons).

Surafel Getachew. 2020. Amharic Facebook Dataset for Hate Speech detection. Mendeley Data.

Fabio Giglietto, Luca Rossi, and Davide Bennato. 2012. The Open Laboratory: Limits and Possibilities of Using Facebook, Twitter, and YouTube as a Research Data Source. *Journal of Technology in Human Services*, 30(3-4):145–159.

Hatebase.org. 2023. Hatebase.org.

Deborah James. 1998. Gender-linked derogatory terms and their use by women and men. *American Speech*, 73(4):399–420.

Akshita Jha and Radhika Mamidi. 2017. When does a compliment become sexist? Analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the 2nd Workshop on NLP and Computational Social Science*, pages 7–16.

Lata Guta Kanessa and Solomon Gizaw Tulu. 2021. Automatic Hate and Offensive speech detection framework from social media: the case of Afaan Oromoo language. In *Proceedings of the 2021 International Conference on Information and Communication Technology for Development for Africa*, pages 42–47. IEEE.

György Kovács, Pedro Alonso, and Rajkumar Saini. 2021. Challenges of hate speech detection in social media: Data scarcity, and leveraging external resources. *SN Computer Science*, 2:1–15.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zelealem Leyew. 2020. Language and society in Ethiopia. *Bulletin of the Department of Linguistics and Philology 40 years*, page 64.

Samuel Minale. 2022. Amharic Social Media Dataset for Hate Speech Detection and Classification in Amharic Text with Deep Learning. Mendeley Data.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text Annotation Tool for Human.

Peace Tech Lab. 2023. Hateful Speech and Conflict in the Federal Democratic Republic of Ethiopia: A lexicon of hateful of inflammatory words and Phrases.

Thomas Poell and José Van Dijck. 2015. Social media and activist communication. *The Routledge companion to alternative and community media*, pages 527–537.

David Shariatmadari. 2016. Eight words that reveal the sexism at the heart of the English language.

Jose Sosa and Serge Sharoff. 2022. Multimodal Pipeline for Collection of Misinformation Data from Telegram. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 1480–1489, Marseille, France. European Language Resources Association.

Ayisha Tabassum and Rajendra R Patil. 2020. A Survey on Text Pre-Processing & Feature Extraction Techniques in Natural Language Processing. *International Research Journal of Engineering and Technology*, 7(06):4864–4867.

Tordue Simon Targema and Joseph M Lucas. 2018. Hate speech in readers' comments and the challenge of democratic consolidation in Nigeria: A critical analysis. *Jurnal Pengajian Media Malaysia*, 20(2):23–38.

Thalikir. 2016. Everyday misogyny: 122 subtly sexist words about women and girls (and what to do about them).

Amsale Zelalem. 2010. *Design and Implementation of Multilanguage Electronic Dictionary for Smart Phones: A Dictionary of Amharic, Afaan Oromo, English and Tigrigna Languages*. Ph.D. thesis, Addis Ababa University.

Low Resource Question Answering: An Amharic Benchmarking Dataset

Tilahun Abedissa Taffa[1,2,3], **Yaregal Assabie**[2], **and Ricardo Usbeck**[3]
[1]Semantic Systems, Universität Hamburg, Hamburg, Germany
[2]Department of Computer Science, Addis Ababa University, Addis Ababa, Ethiopia
[3]Leuphana Universität Lüneburg, Lüneburg, Germany
tilahun.taffa@uni-hamburg.de, yaregal.assabie@aau.edu.et, ricardo.usbeck@leuphana.de

Abstract

Question Answering (QA) systems return concise answers or answer lists based on natural language text, which uses a given context document. Many resources go into curating QA datasets to advance the development of robust QA models. There is a surge in QA datasets for languages such as English; this is different for low-resource languages like Amharic. Indeed, there is no published or publicly available Amharic QA dataset. Hence, to foster further research in low-resource QA, we present the first publicly available benchmarking **Amh**aric **Qu**estion **A**nswering **D**ataset (Amh-QuAD). We crowdsource 2,628 question-answer pairs from over 378 Amharic Wikipedia articles. Using the training set, we fine-tune an XLM-R-based language model and introduce a new reader model. Leveraging our newly fine-tuned reader run a baseline model to spark open-domain Amharic QA research interest. The best-performing baseline QA achieves an F-score of 80.3 and 81.34 in retriever-reader and reading comprehension settings.

Keywords: Low Resource Question Answering, Amharic Question Answering Dataset, Amharic Reading Comprehension, Amh-QuAD

1. Introduction

The task of Question Answering (QA) is to accurately retrieve an answer to a natural language question from a certain underlying data source (Chen and Yih, 2020). The standard train & test QA dataset creation is applied to evaluate models' question synthesis ability and answer accuracy. Crowdsourcing or automatic generation are common approaches in curating QA datasets (Dzendzik et al., 2021). In the crowdsourcing approach, crowd-workers formulate question-answer pairs within a given context. Crowdsourcing allows for the creation of high-quality question-answer pairs, but it is expensive. In contrast, automatic generation approaches leverage language generation models, templates, or machine translation in formulating question-answer pairs. However, attaining a reliable model capable of generating question-answer pairs as accurate as those from a human poses a challenge. Therefore, studies introduce humans in the loop to minimize the generation of trivial, un-grammatical, and incorrect question-answer pairs (Cambazoglu et al., 2020; Fabbri et al., 2020).

The distinction between the existing QA datasets lies in 1) the question expected answer: factoid vs. non-factoid, 2) the data source domain: closed vs. open, and 3) the answer formulation sub-task: extractive vs. generative. Factoid questions like "Who is the founder of Ethiopia's capital Addis Ababa?" (Answer: "Emperor Menelik II") requires a named entity such as proper noun, date, number, or short phrase as an answer (Abedissa and Libsie, 2019).

Context: …በላሊበላ 11 ውቅር ዐብያተ ክርስቲያናት ያሉ ሲሆን ከነዚህም ውስጥ ቤተ ጊዮርጊስ (ባለ መስቀል ቅርፅ) ሊታይ ውሃ ልኩን የጠበቀ ይመስላል። ቤተ መድኃኔዓለም የተባለው ደግሞ ከሁሉም ትልቁ ነው። ላሊበላ (ዳግማዊ ኢየሩሳሌም) የገና በዓል ታህሳስ 29 በልዩ ሁኔታና ድምቀት ይከበራል፣ "ቤዛ ኩሉ" ተብሎ የሚጠራው በነማህ የሚደረገው ዝማሬ በዚሁ በዓል የሚታይ ልዩና ታላቅ ትዕይንት ነው።
(While there are 11 rock-hewn churches in Lalibela, of these churches, *betə giorgis* 'House of St. George' (the one that is cross-shaped) appears to have a leveled foundational platform. The church named *betə medhanialəm* (House of the Saviour of the World), is also the biggest of all. In Lalibela (the Second Jerusalem), *gənna* 'Christmas' holiday is celebrated uniquely and colorfully on December 29. The song called *beza kulu* is played in the aftermath of the holiday and it is a great and special scene observed in this holiday.)
Question: ከላሊበላ አስራ አንዱ ውቅር አብያተ ክርስቲያናት የመስቀል ቅርጽ ያለው የትኛው ነው? (Of the 11 Lalibela's rock-hewn churches, which one is cross-shaped?)
Answer: ቤተ ጊዮርጊስ (*betə giorgis* 'House of St. George')

Figure 1: Amh-QuAD context, question, and answer triplets.

Unlike that, how, why, opinion, definition, and recommendation questions fall into the non-factoid category. For example, a question like "Why does water appear colorless and tasteless?" compels gathering relevant information, reasoning, and synthesizing multiple information pieces from different sources (Yang et al., 2019). Hence, based on the question types, a QA model and its benchmarking dataset are factoid or non-factoid (Dzendzik

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 124–132

et al., 2021). Besides, the data source used to answer a question contains generic information about many things or information specific to a particular domain, like sports, geography, or medicine. Thus, based on the domain of the data source and the question, domain-dependent QA systems are referred to as closed and domain-independent as open QA (Chen and Yih, 2020). Furthermore, QA datasets and models differ in how the answer is retrieved - extractive or generative. Extractive QA datasets like SQuAD (Rajpurkar et al., 2016) measure a QA model competency in predicting the corresponding start and end tokens of the answer span from a context. Unlike that, generative QA datasets contain questions whose answer is a context comprehension, not a direct copy (Raffel et al., 2020).

The architecture of QA systems typically includes question analysis modules to understand questions, information retrieval (IR) systems to locate relevant documents or data and answer extraction mechanisms to extract accurate answers from the retrieved information (Abedissa and Libsie, 2019). In which a natural language question comes into the question analysis module, and an answer flows out of the answer processing module (Chen and Yih, 2020). The question analysis component analyzes the input question in several ways. One is a morphosyntactic analysis, assigning the part-of-speech tag to each word in the question, indicating whether a word is a verb, noun, or adjective. Then, classify questions to identify the semantic type of the question (Utomo et al., 2017). The simplest method of question classification is to use a set of rules that map patterns of questions into question types by analyzing the interrogative terms of the question (wh-terms). However, developing such rules takes time, and adapting to a new domain is challenging. An alternative approach to question classification is the use of machine-learning techniques. This approach treats question type identification using statistical classification packages like a support vector machine (Abedissa and Libsie, 2019). Finally, the question analysis component generates queries from the given question by selecting keywords and removing interrogative terms. In addition, expand the set of keywords using synonyms (Utomo et al., 2017).

The document retrieval component is a standard document retrieval system that identifies a subset of documents that contain terms of a given query from the total document collection deemed most likely to have an answer to the question (Utomo et al., 2017). While trying to identify relevant information more accurately, it splits the documents into several passages and treats the passages as documents. Using a passage-based retrieval approach instead of a full-document retrieval approach has the advantage of returning short text excerpts instead of entire documents, which are easier to process by later components of the question-answering system (Chen et al., 2017).

The answer processing component takes retrieved documents likely to contain an answer to the question and specifies what types of phrases should count as correct answers. Then, it extracts several candidate answers, ranks them in their probable correctness, and returns an answer from those top-ranked phrases. Answer extraction and selection are treated as a classification or ranking problem and solved using heuristics and machine learning methods. Since deep neural networks learn to select features by end-to-end training, most recent QA models use a neural architecture to encode contexts and questions into a vector space and reason over them (Mozannar et al., 2019).

In the era of deep learning, especially pre-trained language models like BERT (Kenton and Toutanova, 2019) enables robust QA model development (Laskar et al., 2020). Besides, the introduction of multi-lingual language models like Cross Language Multilingual-Roberta (XLM-R) (Conneau et al., 2020) and mBERT (Wu and Dredze, 2020) contribute to the advancements of the cross and multi-lingual QA. Existing deep learning-based QA systems fall into retriever-reader, dense retriever and end-to-end training, and retriever-free approaches (Chen and Yih, 2020).

The retriever-reader-based QA models first retrieve relevant passages, then read top-ranked passages and predict the beginning and end positions of the answer text from a context. DrQA (Chen et al., 2017) is a typical example of this approach. In the DrQA model, the retriever uses traditional sparse vector space methods, representing every question and document as bag-of-words vectors weighted by TF-IDF (term frequency-invert document frequency). Then, the retriever passes five top-ranked documents to the reader component. The reader uses a 3-layer bidirectional long-short-term memory (LSTM) (), which encodes the question and the top-ranked paragraphs as a sequence of feature vectors. Then, it predicts the probability of the start and end positions of the answer span (Cui et al., 2019).

The QA models in the retriever-generator approach follow the major paradigm shift towards neural-based IR. To answer a question, generate the response using a retrieved-context instead of predicting start/end positions (Lewis et al., 2020b).

Unlike the retriever-reader and retriever-generator approaches, the generative approach generates free text as an answer to respond to questions using the knowledge in its parameters (Roberts et al., 2020). To test the capability of memorizing factual knowledge of pre-trained language models, Roberts (Roberts et al., 2020)

fine-tuned the T5 (Text-to-Text Transfer Transformer) (Raffel et al., 2020) language model to answer questions without providing it with any additional information or context.

Specific to Amharic, there are very few QA models (Abedissa and Libsie, 2019; Elema, 2022; Yimam and Libsie, 2009); however, none provide a public dataset. Therefore, this paper introduces the first factoid extractive open-domain **Amh**aric **Qu**estion **A**nswering **D**ataset (Amh-QuAD), the dataset can be found online at https://github.com/semantic-systems/amharic-qa.

As shown in Figure Figure 1, the Amh-QuAD dataset comprises context, question, and answer triplets. The contexts consist of articles gathered from Amharic Wikipedia[1], while we crowdsource 2628 question-answer pairs from 378 contexts. For example, for the question given in Figure Figure 1, "ከላሊበላ አስራ አንድ ውቅር አብያተ ክርስቲያናት የመስቀል ቅርጽ ያለው የትኛው ነው?" (Of the 11 Lalibela's rock-hewn churches, which one is cross-shaped?), the answer "ቤተ ጊዮርጊስ" (betə giorgis 'House of St. George') is the span from the context. In our work, in addition to the crowdsourced question-answer pairs, we have set baseline F1-score values by implementing a QA model with the retriever and reader components. We fine-tuned the XLMR model for the reader component using the Amh-QuAD training set and achieved an 81.34 F-score value.

2. Amharic Interrogative Sentences

Amharic, an indigenous African language from Ethiopia, has its unique writing system using the Ge'ez script known as ፊደል (Fidel). As shown in Figure Figure 2, an Amharic interrogative sentence is formulated using information-seeking pronouns like "ምን" (what), "መቼ" (when), "ማን" (who), "የት" (where), "የትኛው" (which) etc. or prepositional interrogative phrases like "ለምን" [ለ-ምን] (why), "በምን" [በ-ምን] (by what), etc. Also, verb phrases such as ግለጽ (explain), ዘርዝር|ሪ (list), አንፃፅር|ሪ|ሩ (compare), etc. are used to pose questions (Yimam, 2009; Amare, 2013).

3. Related Work

Among the existing English QA datasets, SQuAD (Rajpurkar et al., 2016) paved the way by creating question-answer pairs from Wikipedia articles using crowd workers, where each question answer is a span of text in the articles. Chinese MRC (Cui et al., 2019), Vietnamese QA (Do et al., 2021), and other data sets listed in (Dzendzik et al., 2021; Rogers et al., 2023) also follow

[1] https://am.wikipedia.org/

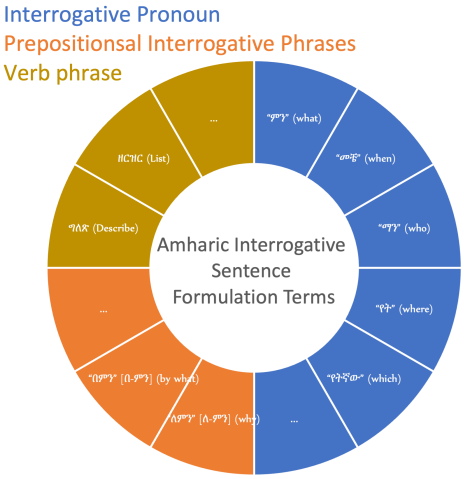


Figure 2: Amharic Interrogative Terms.

the same curation step as SQuAD. Following crowdsourcing, TigQuAD (Gaim et al., 2023) introduces a QA dataset for the low-resourced Semitic language Tigrinya from newspapers. Amharic and Tigrinya are both Semitic languages. However, the linguistic differences in the writing scripts of the two languages (Feleke, 2017) hinder TigQuAD from being used for testing and training Amharic QA models.

On the other hand, by automatically translating SQuAD into their respective languages, German (Möller et al., 2021) and French (d'Hoffschmidt et al., 2020) versions have been created. The Arabic QA dataset (Mozannar et al., 2019) is created partly by translating from SQuAD and partly by crowdsourcing. Translating existing QA datasets to other languages is one solution for creating a large data set. However, we opt for the crowdsourcing approach due to the absence of a well-tested open-source English-to-Amharic machine translation tool.

In Amharic, there are very few QA models; TETEYEQ (Yimam and Libsie, 2009) answers factoid-type questions by extracting entity names using a rule-based answer extraction approach. Abedissa and Libsie(2019) introduce a non-factoid QA model that answers biography, description, and definition questions. The definition-description answer extraction uses heuristics; meanwhile, it answers biography questions using a summarizer and validates the summary with a classifier. The work in (Elema, 2022) classifies questions using a neural network model, selects candidate answers by a hybrid Bi-LSTM and CNN model, and extracts answers as named entities utilizing a named entity recognizer. Unlike the existing Amharic QA systems, this study proposes a retriever-reader-based Amharic QA (AmhQA) that leverages a multi-lingual language model (Conneau et al., 2020). Beyond

attempting to answer Amharic questions, work has yet to produce a published dataset suitable for training and testing the performance of Amharic QA models. Therefore, we present Amh-QuAD as a train & test benchmark for Amharic QA models.

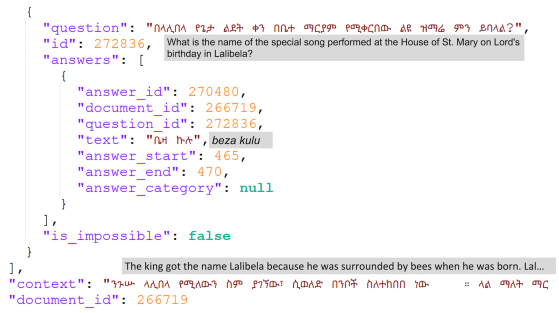


Figure 3: The Amh-QuAD structure.

4. The Amh-QuAD

The Amh-QuAD dataset is created in three phases: article gathering, crowdsourcing question-answer pairs, and annotation.

4.1. Collection and Cleaning

We collect the Amharic articles used as contexts from the Amharic Wikipedia dump[2]. We keep only those articles larger than 2 KB and whose category is not “proverb” and “food preparation”. Proverb articles are advantageous for generating reasoning questions. Additionally, ‘food preparation’ articles mainly consist of steps for preparing food, making them suitable for generating questions such as ‘How is the step to cook...’ and ‘List the steps or ingredients added while cooking...’. Also, in both scenarios, the answer is not confined to a continuous text span within the article but instead spreads out among non-consecutive sentences. We further preprocess the remaining articles after filtration using the wiki-dump-reader tool[3] to obtain clean texts. Subsequently, as long articles do not comprehensively stimulate question creation, each article is segmented based on its sub-topics. Finally, we randomly selected 378 cleaned articles.

4.2. Question-Answer Pair Crowdsourcing

We provide training on formulating questions that can be answered by a given context, following the Haystack guideline[4]. Since we randomly select articles from Wikipedia, we inform annotators to flag any articles containing offensive content. Additionally, we encourage annotators to generate as many questions as possible from a given context.

4.3. Question-Answer Pair Validation and Annotation

The validation of the formulated question-answer pairs is about their correctness. We say a question is correct if the posed question is answerable by the given context, grammatically correct, and clearly defines the subject or object under consideration. For example, a question like ‘How many parks does our (the) country have?’ is ambiguous due to the possessive adjective ‘our’ or the definite article ‘the’; it is challenging to know to which country it refers. We paraphrase such questions according to the context, besides rewriting the questions that do not explicitly state the subject or object. In addition, we exclude questions that are too long and have non-consecutive string answers from the annotation. Then, annotate the question-answer pairs using the Haystack annotation tool[5]. As shown in Figure 3 The annotation tool provides the annotated question-answer pairs as JSON files in SQuAD format.

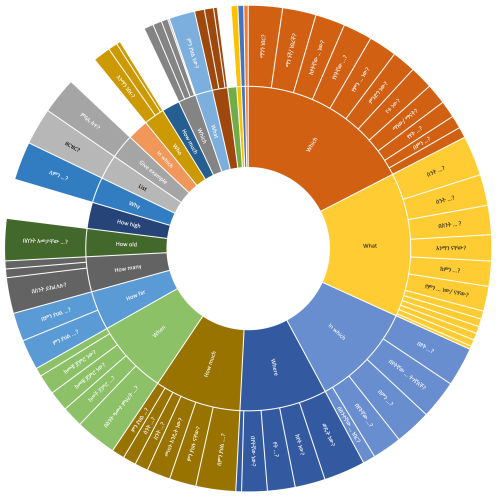


Figure 4: Interrogative terms distribution.

4.4. Data Analysis

As shown in Table Table 1, the Amh-QuAD contains 378 articles and 2628 question-answer pairs. The contexts, on average, have 172 words. Most questions’ average word length is 9.22, whereas the

[2] https://dumps.wikimedia.org/amwiki/20210801/ last accessed 18 August 2021

[3] https://pypi.org/project/wiki-dump-reader/

[4] https://drive.google.com/file/d/1Wv3OIC0Z7ibHIzOm9Xw_r0gjTFmpl-33/view

[5] https://docs.haystack.deepset.ai/docs/annotation

	Article	Question	Answer
Size	378	2628	2628
Word len (avg)	172.07	9.22	2.66

Table 1: Size and average word length of articles, questions, and answers.

answers are short, and their average word length is 2.66. Furthermore, we split the dataset into train, dev, and test with a size of 1728, 600, and 300, respectively. Besides, to see the distribution of interrogative terms on the test set, we manually identify the interrogative term and examine the adjacent two or three tokens. Figure Figure 4 illustrates the distribution of the interrogative terms, showcasing the variety in the question terms within Amh-QuAD.

4.5. Questions Expected Answer Type

Examining the question's interrogative terms and answers, we categorize the 300 test questions into person, location, time, organization, number, description, and other classes. Then, compute the percentage of the coverage of the expected answer types in the test set. As shown in Figure Figure 5, we found that most questions are about Location, Number, and Time, where each type has over 18% coverage. Description questions take 13% of the share and questions that seek a person's name as an answer are 14%. 10% of questions like "What is the working language of Ethiopia?" whose expected answer types are entities that cannot be included in the existing categories and fall into the 'OTHER' group. The list (3%) and organization (3%) are the smallest among the questions.

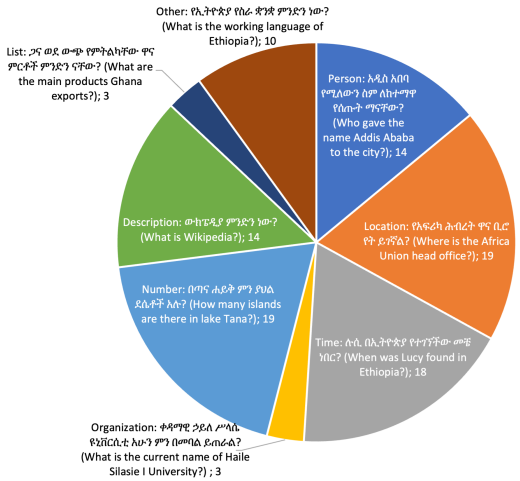


Figure 5: Question Types Distribution in %.

5. Amharic QA Model

Problem Definition: Given a question Q and a set of contexts C, the goal of the Amharic QA (AmhQA) model is to retrieve top-k relevant C_i from C and predict a span of text from the retrieved C_i's that answers Q.

1. Retrieve top-k relevant contexts using a retriever:

$$C_i \in C \quad \text{where} \quad i = 1, 2, \ldots, n$$

2. Predict a span of text from the retrieved context via a reader:

$$S_i = \text{predict}(Q, C_i)$$

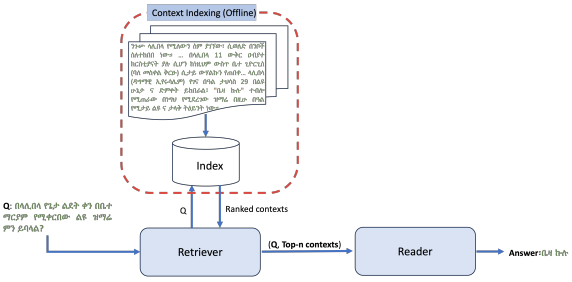


Figure 6: The Amharic QA model.

As shown in Figure 6, the AmhQA model has three components: offline indexer, retriever, and reader.

5.1. Indexing

Offline indexing begins by obtaining contexts from the test dataset. Then, an NLTK-based preprocessor tokenizes the contexts at the word level and splits the contexts into smaller segments based on a maximum length of 200 words with no word overlap. Finally, index the segmented contexts using the Elasticsearch Indexer[6]. This indexing process creates an inverted index, enabling rapid and efficient retrieval of relevant information during subsequent queries.

5.2. Retriever

The AmhQA retriever component utilizes the BM25 (Best Match 25)algorithm to return the top-k most relevant contexts. The BM25 algorithm is a modified TF-IDF(Term Frequency - Inverse Document Frequency) that scores and orders contexts based on their relevance to the given question (Robertson and Zaragoza, 2009). The retriever calculates a

[6] https://www.elastic.co/blog/what-is-an-elasticsearch-index

relevance score for each document by considering term frequencies within documents, document length normalization, and term saturation. Term saturation is the concept that a term's relevance to a document decreases as it appears more frequently within the document. The term saturation function modifies the term frequency during the relevance score calculation.

For a question Q and context C_i, the BM25 scoring formula is:

$$\text{score}(Q, C_i) = \sum_{i=1}^{n} \text{idf}(q_i) \cdot \frac{f(q_i, C_i) \cdot (k_1 + 1)}{f(q_i, C_i) + k_1 \cdot \left(1 - b + b \cdot \frac{|C_i|}{\text{avgcl}}\right)}$$

Where:

- n is the number of terms in the question and q_i is the i-th term in the question.
- $\text{idf}(q_i)$ is the inverse document frequency of term q_i.
- $f(q_i, C_i)$ is the term frequency of term q_i in context C_i.
- $|C_i|$ is the length of context C_i.
- *avgcl* is the average length of contexts in the collection; k_1 and b are tuning parameters.

The parameters k_1 and b control the term frequency component of the scoring. k_1 is a positive tuning parameter that regulates the saturation effect of term frequency. A higher value of k_1 increases the impact of term frequency on the scoring, making the algorithm more sensitive to the frequency of terms in the document. Conversely, a lower value of k_1 reduces the impact of term frequency, leading to less effect on the scoring. The parameter b is a value between 0 and 1 that controls the influence of document length normalization. When b is closer to 0, document length normalization has a weaker effect, resulting in less attenuation of the term frequency component for longer documents. On the other hand, when b is closer to 1, document length normalization has a more substantial effect, causing the term frequency component to favor longer documents.

The retrieved documents are then ranked based on the value of *score(Q, C_i)*.

5.3. Reader

The AmhQA reader component is created by fine-tuning an instance of the XLM-R pre-trained language model from Hugging Face[7] using the open source Haystack framework[8] on our training set. The XLM-R (Cross-lingual Language Model - RoBERTa) (Conneau et al., 2020) is a transformer-based language model trained on diverse languages, including Amharic. While fine-tuning, we use the default settings of the Haystack framework. The reader model generalizes for unseen examples despite being fine-tuned on a small dataset comprising 1728 samples. During the answer retrieval, the reader tokenizes the question Q and context C_i, encodes the tokenized question and context, and produces probability distributions of the answer span start and end indices. Finally, it decodes the answer span indices into human-readable text based on the highest probability span.

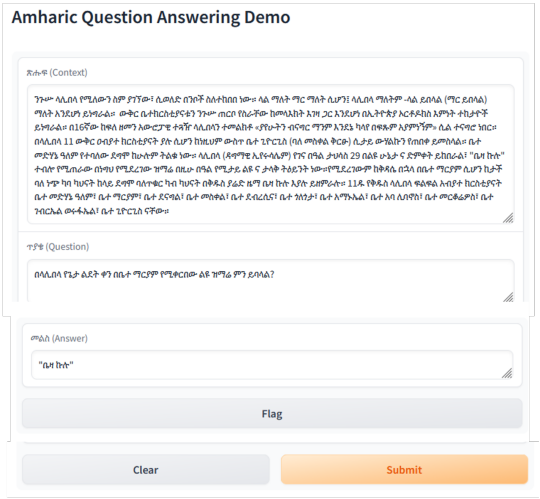


Figure 7: AmhQA Prototype Interface.

6. Experiment

6.1. Baseline Model

Since the Amh-QuAD dataset contains a set of contexts and question-answer pairs, its inherent task is reading comprehension (RC) (Rajpurkar et al., 2016). That is, given a question Q and a context C, the goal of the model is to identify a word or group of consecutive words from C that answers Q. Hence, based on this assumption, we have set a baseline value for the Amh-QuAD using an XLM-R-based RC and with our fine-tuned reader model[9]. Figure 7 shows the RC setting of the AmhQA model prototype interface.

On the other hand, our retriever-reader (RR) based AmhQA model first retrieves relevant passages and then reads top-ranked passages to predict the start and end positions of the answer. The retriever part is based on BM25, and the reader is implemented using our fine-tuned reader model.

[7] https://huggingface.co/deepset/xlm-roberta-large-squad2

[8] https://github.com/deepset-ai/haystack/

[9] https://huggingface.co/deepset/xlm-roberta-large-squad2

Settings	EM	F1
XLM-R on MLQA	52.70	70.7
RC (XLM-R$_{Base}$)	47.49	64.69
RC (XLM-R$_{Large}$)	56.52	74.35
RC (With Fine-tuned Reader)	**67.89**	**81.34**
RR (With Fine-tuned Reader)	67.4	80.3

Table 2: AhmQA performance in RC and RR settings.

6.2. Evaluation and Discussion

The goal of evaluating a QA model is to measure the model's accuracy and its components. For QA datasets where the answer is a span of a text, an exact match (EM) with the gold answer is widely utilized (Rajpurkar et al., 2016). The EM metrics have an all-or-nothing drawback. To overcome this, precision, recall, and their harmonic mean, the F-Score value, is also used (Chen et al., 2017). Recall (R) gives the fraction of words that the system has chosen from the totality of words found in the actual answer, and precision (P) measures the fraction of system answers that are correctly chosen. Besides, Mean Reciprocal Recall (MRR) and Mean Average Precision (MAP) metrics evaluate the retriever performance.

As shown in Table 2, on the RC setting the XLM-R$_{Large}$ F1 score is 74.35, whereas the XLM-R$_{Base}$ F1 score is 64.69. The F1 score of the XLM-R$_{Large}$ on the Amh-QuAD test set was comparable to its average F1 score (70.7) on the MLQA dataset for other seven languages (Lewis et al., 2020a). Our fine-tuned reader also led to substantial improvements, yielding an EM score of 67.89 and an F1 score of 81.34. Even though the difference in the F1 scores achieved by the RC (81.34) and the RR (80.3) settings is minimal, one reason is the segmentation of contexts without overlap during indexing in the RR configuration. The segmentation can split the answer strings into non-overlapping segments, making it difficult for the RR to extract accurate answers. Unlike that, the RC model uses whole context embedding to extract answers from passages, enabling it to achieve better results. Furthermore, the RR includes the retrieval and reading components, introducing complexities in integrating and processing retrieved contexts that affect performance.

6.3. Ablation Study

As shown in Table 3, when the retriever number of context retrieval configuration is top-1, MRR and MAP are high at 82.9, indicating their effectiveness in correctly ranking and retrieving relevant information. Moreover, when expanding the retrieval to the top three results, the scores increase even further. The MRR and MAP reach 88.4 and 88.2, respectively, which indicates that considering multiple retrieval options improves the retriever's ability to capture a broader range of relevant documents, resulting in better ranking and precision. The significant improvement in performance from the top-1 to top-3 settings highlights the necessity of considering multiple retrieval options to optimize the effectiveness of the retriever in the QA models.

	MRR	MAP
top-1	82.9	82.9
top-3	88.4	88.2
top-3**	88.4	88.2

Table 3: AhmQA Retriever component ablation. **(With Fine-tuned Reader)

	EM (top-1)	F1 (top-1)	EM (top-3)	F1 (top-3)
top-1	48.0	60.7	-	-
top-3	53.0	66.6	58.72	73.22
top-3**	50.7	60.9	**67.4**	**80.3**

Table 4: AhmQA Reader component ablation. **(With Fine-tuned Reader)

Table 4 shows the reader component's performance across various metrics and retrieval settings. When considering only the top-1 retrieved context, the Exact Match (EM) and F1 scores are 48.0 and 60.7, respectively. Expanding the retrieved context to the top three results increases the EM and F1 scores at top-1 to 53.0 and 66.6, respectively. Furthermore, when evaluating based on the top three retrieved contexts, both EM and F1 scores experience significant improvements, reaching 58.72 and 73.22, respectively. Highlights the importance of considering multiple retrieved contexts for optimizing the reader's performance, as it allows for a more comprehensive synthesis of contexts.

The fine-tuned reader component has demonstrated a significant performance improvement compared to the previous evaluation. Specifically, the exact match (EM) score has increased to 67.4, indicating higher accuracy in providing precise answers. The F1 score has also improved, reaching 80.3, reflecting enhanced effectiveness in generating correct answers. The top-1 evaluation metrics also show improvements, with exact match top-1 and F1 top-1 scores increasing to 50.7 and 60.9, respectively. These results emphasize the enhanced performance of the fine-tuned reader across different evaluation settings. Overall, the results showcase the improvements achieved through fine-tuning, indicating a more reliable reader component for Amharic QA.

7. Summary

The Amh-QuAD dataset is an effort towards inclusiveness and accessibility in natural language processing (NLP). The development of this dataset will partly address the imbalance in language resources, particularly for underrepresented languages within the NLP community. The Amh-QuAD is the first publicly available factoid open-domain extractive Amharic QA dataset containing triplets of context, question, and answer curated from Amharic Wikipedia, which serves in RC and retriever-reader QA settings. In addition, we introduce a new AmhQA reader by fine-tuning a multilingual pre-trained language model. Also, set baseline values in reading comprehension and retriever-reader QA settings.

8. Bibliographical References

Tilahun Abedissa and Mulugeta Libsie. 2019. Amharic Question Answering for Biography, Definition, and Description Questions. In Fisseha Mekuria, Ethiopia Nigussie, and Tesfa Tegegne, editors, *Information and Communication Technology for Development for Africa*, volume 1026, pages 301–310. Springer International Publishing, Cham. Series Title: Communications in Computer and Information Science.

Getahun Amare. 2013. *Amharic Grammar in Simple Way*. Addis Ababa, Addis Ababa University Press.

B Barla Cambazoglu, Mark Sanderson, Falk Scholer, and Bruce Croft. 2020. A Review of Public Datasets in Question Answering Research. *ACM SIGIR Forum*, 54(2):23.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Danqi Chen and Wen-tau Yih. 2020. Open-Domain Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. 2019. "a span-extraction dataset for Chinese machine reading comprehension". 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)*, pages 5883–5889, Hong Kong, China. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Phong Nguyen-Thuan Do, Nhat Duy Nguyen, Tin Van Huynh, Kiet Van Nguyen, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2021. Sentence Extraction-Based Machine Reading Comprehension for Vietnamese. In *Knowledge Science, Engineering and Management: 14th International Conference, KSEM 2021, Tokyo, Japan, August 14–16, 2021, Proceedings, Part II*, page 511–523, Berlin, Heidelberg. Springer-Verlag.

Daria Dzendzik, Jennifer Foster, and Carl Vogel. 2021. English Machine Reading Comprehension Datasets: A Survey. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8784–8804, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abenezer Mengistu Elema. 2022. Developing Amharic Question Answering Model Over Unstructured Data Source Using Deep Learning Approach. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 108–113.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-Based Question Generation from Retrieved Sentences for Improved Unsupervised Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

Tekabe Legesse Feleke. 2017. The similarity and mutual intelligibility between Amharic and Tigrigna varieties. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 47–54, Valencia, Spain. Association for Computational Linguistics.

Fitsum Gaim, Wonsuk Yang, Hancheol Park, and Jong Park. 2023. Question-answering in a low-resourced language: Benchmark dataset and models for Tigrinya. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11857–11870, Toronto, Canada. Association for Computational Linguistics.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Md Tahmid Rahman Laskar, Jimmy Xiangji Huang, and Enamul Hoque. 2020. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5505–5514, Marseille, France. European Language Resources Association.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020a. MLQA: Evaluating Cross-lingual Extractive Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Timo Möller, Julian Risch, and Malte Pietsch. 2021. GermanQuAD and GermanDPR: Improving non-English question answering and passage retrieval. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 42–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv:1606.05250 [cs]*. ArXiv: 1606.05250.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Comput. Surv.*, 55(10).

Fandy Setyo Utomo, Nanna Suryana, and Mohd Sanusi Aami. 2017. Question Answering System: A Review on Question Analysis, Document Processing, and Answer Extraction Techniques. *Journal of Theoretical & Applied Information Technology*, 95(14).

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. Association for Computational Linguistics.

Baye Yimam. 2009. *Amharic Grammar*. Addis Ababa, Addis Ababa University Press.

Seid Muhie Yimam and Mulugeta Libsie. 2009. TETEYEQ: Amharic question answering for factoid questions. *IE-IR-LRL*, 3(4):17.

The Annotators Agree To Not Agree On The Fine-grained Annotation of Hate-speech against Women in Algerian Dialect Comments

Imane Guellil[1], **Yousra Houichi**[2], **Sara Chennoufi**[3], **Mohamed Boubred**[4], **Anfal Yousra Boucetta**[5], **Faical Azouaou**[6]
[1]University of Edinburgh, Edinburgh, United Kingdom
[2]EZUS, France,
[3]LIRIS, France,
[4]Capgemini, France
[5]Ecole Superieure d'Informatique D'Alger (ESI), Algeria,
[6]Ecole supérieure en Sciences et Technologies de l'Informatique et du Numérique (ESTIN), Algeria

Abstract

A significant number of research studies have been presented for detecting hate speech in social media during the last few years. However, the majority of these studies are in English. Only a few studies focus on Arabic and its dialects (especially the Algerian dialect) with a smaller number of them targeting sexism detection (or hate speech against women). Even the works that have been proposed on Arabic sexism detection consider two classes only (hateful and non-hateful), and three classes(adding the neutral class) in the best scenario. This paper aims to propose the first fine-grained corpus focusing on 13 classes. However, given the challenges related to hate speech and fine-grained annotation, the Kappa metric is relatively low among the annotators (i.e. 35%). This work in progress proposes three main contributions: 1) Annotation of different categories related to hate speech such as insults, vulgar words or hate in general. 2) Annotation of 10,000 comments, in Arabic and Algerian dialects, automatically extracted from Youtube. 3) Highlighting the challenges related to manual annotation such as subjectivity, risk of bias, lack of annotation guidelines, etc.

Keywords: Sexism detection, hate-speech detection, corpus construction, manual annotation

1. Introduction

Hate speech is commonly defined as a language to express hatred against a specific person or a group based on certain key characteristics such as religion, gender, race, sexual orientation, and various disability forms (Shannaq et al., 2022). The excessive use of social media leads to the rise of antisocial behaviours illustrated in the spread of online hate speech, offensive language and cyberbullying (Shannaq et al., 2022). Authorities in many countries are recognizing hate speech as a serious problem as it can lead to depression which hurts people's health and relationships. It can also lead to suicide in more serious scenarios (Boucherit and Abainia, 2022).

With the online proliferation of hate speech, a significant number of research studies focusing on how to classify and detect this kind of speech have been presented in the last few years. The majority of these studies detect general hate speech (Caiani et al., 2021; Pamungkas et al., 2018; Almatarneh et al., 2019; Kalaivani and Thenmozhi, 2021) and only a few studies (de Paula et al., 2021) focused on the detection of hate speech against women (only by distinguishing between hateful and non-hateful comments). However, almost all studies are dedicated to English. This is mainly due to the lack of resources (lexicons and corpora that are constructed for other languages such as Arabic). To bridge the gap, the role of this paper is to propose a fine-grained manually annotated corpus including 10,000 YouTube comments and 13 classes: 0 (no hate), i (insult), v (vulgar), h (hate), s (without relationships with women), b (positive), p (a problem in the annotation), e (emojis only), c (passage from Coran, Muslims book), iv (insults and vulgar in the same time), ih (insult and hate in the same time), vh (vulgar and hate in the same time), ivh (insult, vulgar and hate in the same time). This corpus will be freely available to the research after its publication. The main conclusion from this work was that annotators tend to disagree more frequently when they have to deal with different annotation classes.

2. Arabic Hate Speech in social media: Challenges

Arabic is a language spoken by more than 330 million people as a native language. It is the fifth most spoken language in the world. Modern Standard Arabic (MSA) is usually the official language used in school whereas the classical is used in the Holy Qur'an (Muslim's book) (ESI, 2016; Guellil et al., 2020b). Another form of Arabic is the Arabic di-

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 133–139

alects which are used in daily life conversations. Also, Arabic in social media can be written either by using Arabic letters or Arabizi (Latin letters) (Guellil et al., 2021). 55% of the text in social media was written in Arabic (2017) (Haddad et al., 2020). Arabic Natural Language Processing (NLP) applications have to deal with several complex challenges in addition to the common challenges related to any NLP problems(Guellil and Faical, 2017; Guellil et al., 2018).

Arabic is known for its challenges, scarcity of resources and complexity. Detecting hate speech for Arabic content is a complex task (Husain, 2020). Different challenges can be raised when detecting hate speech in Arabic text: 1) The informal language using short forms and slang. 2) The use of dialects (Boucherit and Abainia, 2022). 3) The diversity of the Arabic language dialects (Husain, 2020). 4) The use of Arabizi (Guellil et al., 2020a)

3. Related Work on Arabic hate-speech corpora creation

Some papers focused on resources constructions dedicated to hate-speech detection (Albadi et al., 2018; Mubarak et al., 2022; Alsafari et al., 2020; Mubarak et al., 2020; Chowdhury et al., 2020; Almanea and Poesio; El Abboubi et al., 2020; Boucherit and Abainia, 2022; Guellil et al., 2022). (Albadi et al., 2018) aims to detect religious hate speech in the Arabic language on social media[1]. The authors started with constructing their dataset by collecting tweets and annotating them manually. For this purpose, They first collected 6,000 Arabic tweets referring to different religious groups and labelled them using crowdsourced workers. After this, they analysed the labelled dataset and reported the main targets of religious hatred in the Arabic Twitter space.

In the paper of Mubarak et al. (2022), the authors present an automated emoji-based approach of collecting tweets that have a much higher percentage of malicious content, without having any language dependency. From a collection of 4.4M Arabic tweets between June 2016 and November 2017, they extracted all tweets having any of the used emojis. An annotation job was created on the Appen crowdsourcing platform to judge whether a tweet is offensive or not. Annotators from all Arab countries were invited.

The role of the paper described by Alsafari et al. (2020) was to create a reliable Arabic textual corpus. The Data was extracted from Twitter based on a list of Arabic keywords related to each of the four categories under study: religion, ethnicity, nationality and gender. The authors randomly selected 200,000 posts for each category, with a total of 800,000 samples. The annotation has been carried out by three Gulf native speakers, two females and one male.

The paper of Mubarak et al. (2020) is adding an additional class to those which are generally studied, where these authors also identify vulgar comments in addition to comments including hate. The Twitter APIs were used to collect 660k Arabic tweets between April 15 – May 6, 2019. The tweets were annotated, ending up with 1,915 offensive tweets. Each tweet was labelled as offensive, which could additionally be labelled as vulgar and/or hate speech, or Clean.

The main idea of Chowdhury et al. (2020) was to introduce a new dialectal Arabic news comment dataset, collected from multiple social media platforms, including Twitter, Facebook, and YouTube. From 2011 to 2019, over 100k comments from different social media platforms were collected. The contents from each platform were collected through its own API (YouTube, Facebook, and Twitter). Data annotation (Amazon Mechanical Turk (AMT), a crowdsourcing platform, was used to obtain manual annotations. The comments were annotated for hate speech and vulgar (but not hate) categories. The authors analyzed the distinctive lexical content along with the use of emojis in offensive comments.

The aim of Almanea and Poesio was to introduce an Arabic misogyny and sexism dataset (ArMIS) characterized by providing annotations from annotators with different degrees of religious beliefs and providing evidence that such differences do result in disagreements. The authors discussed proof-of-concept experiments showing that a dataset in which disagreements have not been reconciled can be used to train state-of-the-art models for misogyny and sexism detection; and considered different ways in which such models could be evaluated.

The aim of El Abboubi et al. (2020) was to discuss both the impact of possible sex-based differences and the awareness and recognition of sexist attitudes in Moroccan Arabic. The findings of this study are based on quantitative data. The patterns analyzed are the following: sexist attitudes, self-assessment, sources of pressure to use or change sexist language, and recognition of sexist language. A questionnaire was designed to measure attitudes. The questionnaire is divided into two parts: one in which five questions are asked to reflect the respondents' attitudes towards Moroccan Arabic as a sexist language; and a second part in which statements are presented to respondents who rate them considering the extent to which they are sex-

[1] https://github.com/nuhaalbadi/Arabic_hatespeech

ist, and if those same statements are appropriate or not.

The paper of Boucherit and Abainia (2022) addresses the problem of detecting offensive and abusive content in Facebook comments, where the focus is on the Algerian dialectal Arabic. The authors have built a new corpus regrouping more than 8.7k texts manually annotated as normal, abusive and offensive (where 10,258 comments have been initially collected from public pages and groups related to sensitive topics).

In the paper of Mulki et al. (2019), the authors introduced the Levantine Hate Speech and Abusive (L-HSAB) Twitter dataset to be a benchmark dataset for automatic detection of online Levantine toxic contents. Three annotators manually labelled the tweets following into 3 categories: Normal, Abusive and Hate. Waseem et al. (Waseem and Hovy, 2016) manually annotated the dataset containing 16,914 tweets where 3,383 tweets were for sexist content, 1,972 for racist content, and 11,559 for neither sexist nor racist. For dataset generation, the authors used Twitter API to extract tweets containing some keywords related to women. The work of Waseem et al. (Waseem and Hovy, 2016) is considered a benchmark by many researchers (Al-Hassan and Al-Dossari, 2019; Pitsilis et al., 2018; Kshirsagar et al., 2018).

Finally, our recent work Guellil et al. (2022), also considered YouTube for constructing a corpus of 5,000 comments dedicated to sexism detection. However, we only considered two labels for annotating their dataset: Hateful and non-hateful comments.

4. Data collection and annotation

4.1. Data collection

Youtube comments related to videos about women are used. A feminine adjective such as: جميلة meaning beautiful, جَايحة meaning stupid or كلبة meaning a dog are targeted. A video on YouTube is recognised by a unique identifier (*video_id*). For example, the video having an id equal to *"TJ2WfhfbvZA"* handling a radio emission about unfaithful women and the video having an id equal to *"_VimCUVXwaQ"* advises women to become beautiful. Three annotators manually reviewed the obtained video from the keyword and manually selected 335 *video_id*. We used Youtube Data API[2] and a Python script to automatically extract comments of each *video_id* and their replies. In the end, we were able to collect *373,984* comments extracted, we call this corpus *Corpus_Youtube_women_10000*.

[2]https://developers.google.com/youtube/v3/

4.2. Data annotation

For the annotation, we randomly select 10,000 comments containing MSA and Algerian dialects written in Arabic and Arabizi. This corpus also contains some comments in French and others in English (As most of the Arabic people are bilingual). The annotation was done by three annotators, native speakers of Arabic and its dialects (2 women and one man). The annotators were separated and they had 3 weeks to manually annotate the selected comments using different labels. An annotation guideline was prepared for this purpose and it was shared with the annotators. The main points of this guideline are:

- The value of the column *hate* can be given multiple values: 0 (comment containing no hate, no insult, no vulgar word), i (if the comment contains insults, for example, *ya kalba meaning dog*, *ya hmara meaning donkey*, etc,), v (if the comment contains vulgar words), h (if the comment contains hate, for example *allah ya3tik elmoutI want meaning that you die*, or *we will dance on your grave, etc)*.
- If it has a comment that contains several characteristics at the same time, they had to mention it. For example, if a comment contains hate and vulgarities, you had to put vh (and not hv), in the same order i, v then h- had to be kept.
- The authors were asked to be as objective as possible for this annotation and not incorporate their personal feelings.
- As the comments were extracted automatically, it is also possible to find some comments with no relationship with women As an example *four lghounia hadi kho meaning this song is amazing bro*, They were asked to put the letter s (without interest)
- They were asked to put the letter p (problem) When they were facing a situation where they could not decide what to put. However, they were asked to use this option only when it is necessary.
- As we plan to use this corpus for sentiment analysis purposes as well, the annotators were also asked to put a b for the positive comments.
- The comments including only emojis without text should be annotated with the label e
- The comments including punctuation only should be annotated with the word "po"

Table 1: Agreement of the three annotators on the different labels

Label	rater1	rater2	rater3	Agreement
0	6723	3310	1206	695
i	1109	749	1843	285
v	266	366	338	107
h	106	1012	128	41
s	222	1583	3912	121
b	1027	2199	1809	869
p	103	8	129	0
e	82	25	0	0
c	8	1	8	0
iv	138	72	219	9
ih	115	402	151	27
vh	7	70	25	1
ivh	12	131	40	2

- The comments including only some text from the Coran should be annotated as c

4.3. The constructed corpus

The table illustrates the number of labels given by each annotator. Rater1 and Rater2 are women. Rater3 is a man.

Table 1 illustrates the agreement of the three annotators on the different labels. From this table, we observe that the annotators tend to more agree on the positive reference or the non-hateful references than they agree on the hateful comments. We also observe that the disagreement is higher for comments including more than one hateful class (such as comments including insults and vulgar words simultaneously). Finally, we can also observe that three tendencies of annotations are among our annotators. We have the careful annotation (Rater1) when the annotator does not assign a label only when she is sure. We have the extreme annotation (Rater2) when the annotators assigned the majority of the labels and we have the moderate annotator (Rater3) who tends to be in between the two previous annotators.

In order to highlight the inter-agreement among annotators we also present Figure 1 illustrating the Kappa-agreement between each two annotators. We observe that this rate is especially low between rater1 and rater3 (19%). The best agreement is between Rater2 and Rater3. We observe that fine-grain annotation with many classes returns a low kappa (illustrated in Figure 1). One of the reasons behind this is the typing errors related to some labels. This is also caused by the non-application of the guideline. For example, one of the annotators created another label ("other") when he should have used the label p for the problems. Another cause of conflicts is when the authors have to attribute different labels to the same comments. We observe a lack of consistency where some annotators misplaced the labels.

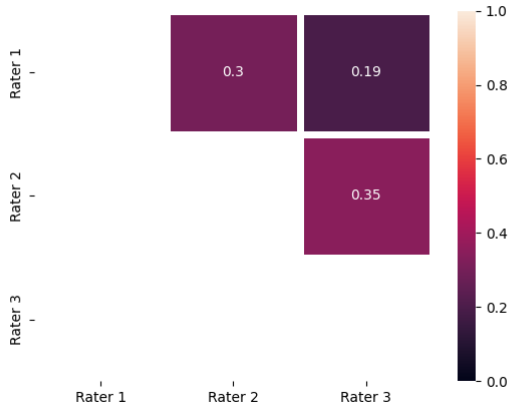


Figure 1: Intra-agreement among annotators

5. Discussion

In total 10,000 comments that were randomly selected were annotated by three annotators. However, we can observe that the inter-agreement among annotators (Kappa) was really low. This highlights how complicated is the annotation with many labels. In total, the 3 annotators agree on 2,157 (22%) comments from the 10,000 that they initially reviewed.

The main goal of this paper is to propose a resource for fine-grained hate speech detection. However, this resource can also be used for binary classification (when the research aims to only detect hate speech against women). In order to do that, we need to first separate the labels into two categories to distinguish between hateful and not-hateful comments. We decide to recognise the labels 0, s, b, p, e, c as non-hateful and the others (i, v, h, iv, ih, vh, ivh) as hateful. We also resolve some obvious annotation errors such as the one related to the tag "other" that we recognise as non-hateful. In that case, we observe that the three annotators agree on 1165 hateful comments and on 6219 non-hateful comments (a total of 7384 comments). The intra-agreement among annotators is illustrated in Figure 2. We observe in this figure that Kappa significantly improves, especially between the second and the third annotators where Kappa with two classes is up to 0.68 (considered to be a good degree of agreement (Salkind, 2010)). Hence, in all cases, we observe that Rater2 is providing the highest agreement.

The main challenge when annotating a corpus with many labels is the consistency of annotation guide-

lines. The annotators have different questions at the start of the annotation phase. The best way to do this would be to have an annotation pilot by selecting only a few documents (around 20) having them annotated by the three annotators and having a discussion for resolving the disagreements before starting the annotation. Another issue is the lack of consistency among the annotators. Some annotators created new classes when others did not respect the annotation format. One way to resolve this would be to automatically detect this incoherence and have it reviewed manually again by the annotators.

This corpus may be used in different ways. The first one would be to train a binary classifier for detecting hateful and not hateful comments. We can observe that the agreement for the binary classification is pretty good. However, the main aim of this corpus is to train a multi-class classifier in order to automatically distinguish among hate, insult and vulgar comments used against women in social media. The main challenge behind this would be the imbalance of the different classes. We can consider the augmentation of some classes. We can also consider algorithms dedicated to handling imbalanced corpora such as the Synthetic Minority Oversampling Technique (SMOTE)

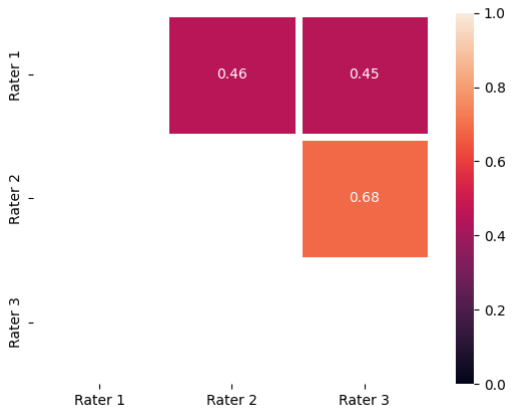


Figure 2: Intra-agreement among annotators

6. Conclusion

We constructed in this paper the first fine-grained corpus for Arabic/Algerian dialect hate speech against women detection. We focus on Arabic/Algerian dialect but we plan to extend this construction to other dialects such as Morrocan or Tunisian. We plan to extend this construction to other African languages as well. This corpus includes 14 labels and is distinguished among the general hate, insults and vulgar comments. Our future would be to automatically review some disagreements related to the mismatch of labels, upper-case, etc. We also plan to have this annotation reviewed by a fourth annotator who will have access to the different assigned labels in addition to the comments. We also plan to use the constructed corpus in order to train ML algorithms for fine-grained classification.

Acknowledgement

We would like to thank the Edinburgh Futures Institute (EFI)[3] for supporting the fees related to paper presentations and research trips leading to such quality papers. The purpose of the Edinburgh Futures Institute (EFI) is to pursue knowledge and understanding that supports the navigation of complex futures.

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: A survey on multilingual corpus. *Computer Science & Information Technology (CS & IT)*, 9(2):83.

Albadi, Nuha and Kurdi, Maram and Mishra, Shivakant. 2018. *Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere*. IEEE.

Dina Almanea and Massimo Poesio. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements.

Sattam Almatarneh, Pablo Gamallo, Francisco J Ribadas Pena, and Alexey Alexeev. 2019. Supervised classifiers to identify hate speech on english and spanish tweets. In *International Conference on Asian Digital Libraries*, pages 23–30. Springer.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. 2020. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096.

Oussama Boucherit and Kheireddine Abainia. 2022. Offensive language detection in under-resourced algerian dialectal arabic language. *arXiv preprint arXiv:2203.10024*.

Manuela Caiani, Benedetta Carlotti, and Enrico Padoan. 2021. Online hate speech and the radical right in times of pandemic: The italian and english cases. *Javnost-The Public*, 28(2):202–218.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J

[3]https://efi.ed.ac.uk/

Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the 12th language resources and evaluation conference*, pages 6203–6212.

Angel Felipe Magnossão de Paula, Roberto Fray da Silva, and Ipek Baris Schlicht. 2021. Sexism prediction in spanish and english tweets using monolingual and multilingual bert and ensemble models. *arXiv preprint arXiv:2111.04551*.

Zineb El Abboubi, Ahmadou Bouylmani, and Mohammed Derdar. 2020. Sexism in moroccan arabic: Gender differences in perceptions and use of language. *Journal of Applied Language and Culture Studies*, 3:215–230.

Karima Benatchba Prof President ESI. 2016. *A Sentiment analysis approach for Arabic dialects texts analysis based on automatic translation: Application to the Algerian dialect.* Ph.D. thesis, ESI Algeria.

Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi, and Akram Abdelhaq Moumna. 2022. Ara-women-hate: An annotated corpus dedicated to hate speech detection against women in the arabic community. page 68–75.

Imane Guellil, Faical Azouaou, Fodil Benali, Ala Eddine Hachani, and Marcelo Mendoza. 2020a. The role of transliteration in the process of arabizi translation/sentiment analysis. *Recent Advances in NLP: The Case of Arabic Language*, pages 101–128.

Imane Guellil, Faical Azouaou, Fodil Benali, alaeddine Hachani, and Houda Saadane. 2018. Approche hybride pour la translitération de l'arabizi algérien : une étude préliminaire. In *Conference: 25e conférence sur le Traitement Automatique des Langues Naturelles (TALN), May 2018, Rennes, FranceAt: Rennes, France*. https://www.researchgate.net/publication ....

Imane Guellil, Faical Azouaou, and Francisco Chiclana. 2020b. Arautosenti: automatic annotation and new tendencies for sentiment classification of arabic messages. *Social Network Analysis and Mining*, 10:1–20.

Imane Guellil and Azouaou Faical. 2017. Bilingual lexicon for algerian arabic dialect treatment in social media. In *WiNLP: Women & Underrepresented Minorities in Natural Language Processing (co-located with ACL 2017)*.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507.

Bushr Haddad, Zoher Orabe, Anas Al-Abood, and Nada Ghneim. 2020. Arabic offensive language detection with attention-based deep neural networks. In *Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection*, pages 76–81.

Fatemah Husain. 2020. Arabic offensive language detection using machine learning and ensemble machine learning approaches. *arXiv preprint arXiv:2005.08946*.

Adaikkan Kalaivani and Durairaj Thenmozhi. 2021. Multilingual hate speech and offensive language detection in english, hindi, and marathi languages.

Rohan Kshirsagar, Tyus Cukuvac, Kathleen McKeown, and Susan McGregor. 2018. Predictive embeddings for hate speech detection on twitter. *arXiv preprint arXiv:1809.10644*.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. 2020. Arabic offensive language on twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Hala Mulki, Hatem Haddad, Chedi Bechikh Ali, and Halima Alshabani. 2019. L-hsab: A levantine twitter dataset for hate speech and abusive language. In *Proceedings of the third workshop on abusive language online*, pages 111–118.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. Automatic identification of misogyny in english and italian tweets at evalita 2018 with a multilingual hate lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, volume 2263, pages 1–6. CEUR-WS.

Georgios K Pitsilis, Heri Ramampiaro, and Helge Langseth. 2018. Detecting offensive language in tweets using deep learning. *arXiv preprint arXiv:1801.04433*.

Neil Salkind. 2010. Cohen's kappa. pages 188–189.

Fatima Shannaq, Bassam Hammo, Hossam Faris, and Pedro A Castillo-Valdivieso. 2022. Offensive language detection in arabic social

networks using evolutionary-based classifiers learned from fine-tuned embeddings. *IEEE Access*, 10:75018–75039.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Advancing Language Diversity and Inclusion: Towards a Neural Network-based Spell Checker and Correction for Wolof

Thierno Ibrahima Cissé, Fatiha Sadat
Université du Québec à Montréal
Montréal, Canada
cisse.thierno_ibrahima@courrier.uqam.ca, sadat.fatiha@uqam.ca

Abstract

This paper introduces a novel approach to spell checking and correction for low-resource and under-represented languages, with a specific focus on an African language, Wolof. By leveraging the capabilities of transformer models and neural networks, we propose an efficient and practical system capable of correcting typos and improving text quality. Our proposed technique involves training a transformer model on a parallel corpus consisting of misspelled sentences and their correctly spelled counterparts, generated using a semi-automatic method. As we fine tune the model to transform misspelled text into accurate sentences, we demonstrate the immense potential of this approach to overcome the challenges faced by resource-scarce and under-represented languages in the realm of spell checking and correction. Our experimental results and evaluations exhibit promising outcomes, offering valuable insights that contribute to the ongoing endeavors aimed at enriching linguistic diversity and inclusion and thus improving digital communication accessibility for languages grappling with scarcity of resources and under-representation in the digital landscape.

Keywords: Spell check and correction, low-ressource language, Wolof, endangererd, Indigenous, parallel corpus, Transformer.

1. Introduction

In recent years, Natural Language Processing (NLP) has made impressive progress in understanding, analyzing and generating human language. Yet, most of this progress is focused on high-resource languages like English, French, and Spanish, leaving low-resource and under-represented languages with limited tools and resources for effective NLP applications. This paper aims to bridge this gap by introducing a novel approach for spell checking and correction in resource-scarce languages. Specifically, we focus on Wolof, an African spoken language that has recently sparked interest in NLP research. We also present a new dataset that can be used for word correction in Wolof. This study contributes to the overarching objective of developing inclusive and effective natural language processing (NLP) tools and resources, in alignment with the ethos of ”no language left behind”.

Wolof, a Senegambian language primarily spoken in Senegal, Gambia and Mauritania (Diouf et al., 2017), serves as an example of a low-resource language that could greatly benefit from NLP advancements. Despite having over 10 million native speakers (Eberhard et al., 2019), there is a significant lack of digital resources and computational tools for most of (if not all) African languages, among them the Wolof language. As the world increasingly connects through digital platforms, it is vital to ensure robust NLP tools are available for low-resource languages like Wolof. Providing speakers of the language with accurate and effective spell checking and correction systems can enhance linguistic accessibility and promote digital communication across diverse linguistic communities.

Developing spell checkers and correction systems for low-resource languages is difficult due to the limited availability of annotated data, morphological complexity, and the absence of well-established computational resources. Traditional methods like rule-based or dictionary-based systems may not adequately address these challenges, requiring alternative approaches. Deep learning techniques, particularly transformer models, have demonstrated immense potential in various NLP tasks lately. These techniques can learn complex language patterns and generate context-sensitive representations, making them ideal for tackling challenges associated with low-resource language spell checking and correction.

This paper presents a transformer-based model for word correction and spelling in Wolof. Our model is trained on a parallel corpus consisting of misspelled sentences and their error-free counterparts, optimizing the model to translate error-prone text into accurate sentences. Furthermore, we contribute to the advancement of NLP for the Wolof language by creating a new corpus of misspelled sentences and their error-free counterparts. This corpus serves as a benchmark and state-of-the-art in word correction and spelling for Wolof, provid-

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 140–151

ing a valuable resource for future research. This resource will facilitate the development of more advanced NLP tools and applications for Wolof.

The remainder of this paper is organized as follows: Section 2 reviews related work on the Wolof language and offers an overview of low-resource language spell checking and correction, as well as neural networks and transformer models in NLP. Section 3 details the methodology employed in developing our transformer-based spell checking and correction system. Section 4 presents our evaluation results, including a discussion of the system's performance. In Section 5, we examine the limitations of our system and discuss potential areas for improvement. Finally, Section 6 concludes the paper, underlines the implications of our findings and suggests future research directions.

2. Background

2.1. Wolof Language

Wolof is a language belonging to the Senegambian group within the Northern branch of the Atlantic language family, which is part of the broader Niger-Congo language family. It shares strong linguistic connections with Pulaar and Serer languages (Sapir, 1971; Doneux, 1978; Wilson, 1989). The Atlantic language family includes approximately 40 languages, with Pulaar (a dialect of Fula) being the exception, and most are spoken in regions near of the Atlantic coast of Africa. Although Wolof is fundamentally an oral language, its orthography was standardized in 1972 (Robert, 2011).

Descriptive linguistic studies of Wolof can be traced back to the colonial period (Boilat, 1858), while other researches on Wolof morphology and syntax have been conducted by Diagne (1971), Mangold (1977), Church (1981), Dialo (1981), and Ka (1981). In-depth analytical studies of Wolof syntax can be primarily found in the works of Njie (1982) and Dunigan (1994).

Wolof is mainly an aspectual language, focusing on the aspect of an action rather than its tense. This characteristic allows the imperfective marker to combine with various tense markers. The language features a rich verb system, which includes a wide array of basic verbal forms and paradigms. Notably, Mangold (1977) and Church (1981) provide systematic presentations of Wolof verbal paradigms.

In terms of literature and resources, Wolof appears in various forms, such as novels, short story collections, and poetry. However, even in Senegal, it is challenging to find materials written in Wolof. Recent efforts have been made to improve the availability of resources for Wolof speakers. In a study by Gauthier et al. (2016), researchers gathered an Automatic Speech Recognition (ASR) dataset for four African languages, including Wolof. This dataset was then used to create the first ASR system for Wolof. Another initiative was proposed by Nguer et al. (2015), who outlined the creation process for the first collaborative online Wolof dictionary. This project was part of the larger Dictionnaires Langues Africaines - Français (DiLAF) project[1], which has produced dictionaries for seven African languages, including Wolof. However, at the time of writing, all dictionaries are accessible online except for the Wolof one. More recently, Cissé and Sadat (2023) have presented a range of resources for the Wolof language, including a spell checking tool mainly grounded in the language's writing rules.

2.2. Low-Resource Language Spell Checking and Correction

Spell checking and correction for low-resource languages have been of great interest to many researchers. Early approaches often depended on rule-based systems (Armstrong et al., 1995) or statistical methods, such as noisy channel models (Kernighan et al., 1990), n-gram models (Stolcke, 2000), and hidden Markov models (Viterbi, 1967). However, these methods often require substantial linguistic knowledge and annotated data, which may be scarce or non-existent for low-resource languages.

In recent years, researchers have investigated data-driven approaches for low-resource languages, such as unsupervised learning (Soricut and Och, 2015) and bootstrapping techniques (Yarowsky et al., 2001). Some studies have also explored the use of cross-lingual transfer learning (Täckström et al., 2012) or leveraging comparable corpora (Madnani et al., 2012) to enhance spell checking and correction performance in low-resource languages. Nevertheless, these approaches may still be constrained by the availability and quality of parallel and comparable corpora.

2.3. Neural Networks in Spell Checking and Correction

The emergence of deep learning techniques, in particular transformer models (Vaswani et al., 2017) and neural networks, has had a significant impact on the NLP field. These techniques have shown immense potential in a wide range of tasks, including machine translation (Bahdanau et al., 2015), information retrieval, conversational agents, sentiment analysis (Socher et al., 2013), and text summarization (See et al., 2017).

[1] http://pagesperso.ls2n.fr/~enguehard-c/DiLAF/index.php

In the context of spell checking and correction, sequence-to-sequence models (Sutskever et al., 2014) have been employed with promising results, using an encoder-decoder architecture to map misspelled sequences to their correct counterparts (Hládek et al., 2019). Attention mechanisms (Bahdanau et al., 2015) have also been integrated into these models to enhance the alignment between input and output sequences (Garg et al., 2019).

The development of transformer models has further advanced the capabilities of neural networks in spell checking and correction. Transformer models, which rely on self-attention mechanisms, have proven effective in capturing long-range dependencies and providing more accurate context-sensitive representations (Devlin et al., 2019). Recent studies have applied transformer models, such as BERT and GPT (Radford and Narasimhan, 2018), to spelling error detection and correction (Sorokin et al., 2016), or fine-tuned them for specific low-resource languages (Al-Ghamdi et al., 2023).

3. Methodology

Our approach consists of three main steps, namely data preparation, model architecture building, and model training configuration.
Initially, we discuss the process of data acquisition and corpus annotation, which is crucial for training an effective model, especially in the context of low-resource languages. Subsequently, we delve into the architecture of the transformer model, detailing its components and design choices. Finally, we describe the training configurations, including the parameters and settings used to train the model.

3.1. Data selection and annotation process

The data acquisition and corpus annotation process encompasses two principal phases. Initially, we identified suitable sources for the corpus data, which were available in various formats (e.g., PDF, text, HTML), and subsequently carried out the extraction of content. Following this, we employed a hybrid approach, incorporating both manual and automatic annotation techniques, and conducted thorough proofreading to generate a corpus of accurately corrected sentences.

3.1.1. Data Selection

The data collection process for our Wolof spell correction study involved gathering data from various sources such as news websites[2], social media platforms[3] [4], religious websites[5], religious PDF files (Diagne, 1997), bilingual Wolof-French dictionaries (Diouf and Kenkyūjo, 2001; Cissé, 2004) and bilingual Wolof-French corpora released (Adelani et al., 2022; Costa-jussà et al., 2022).
In total, we collected 78,384 sentences for our corpus. During the collection process, we emphasized the quality and diversity of the content, ensuring that our corpus included sentences from various domains and genres.

3.1.2. Corpus annotation

First, we used Python scripts to scrape data from news websites, social media platforms, and religious websites. This process yielded 25,860 sentences from religious websites, 21,341 sentences from social media platforms, and 13,245 sentences from news websites. Next, we extracted 10,087 sentences from religious PDF files and Wolof-French bilingual dictionaries. Additionally, we used the Wolof data from the bilingual Wolof-French corpora released by Masakhane (Adelani et al., 2022) and Facebook (Costa-jussà et al., 2022; Goyal et al., 2022). The detailed statistics of each corpus, including the number of sentences, are outlined in Table 1.

Splits	Masakhane	Facebook
Train	3360	997
Dev	1506	1012
Test	1500	N/A

Table 1: Corpora statistics

All collected sentences were saved in plain text files using the UTF-8 encoding. We observed that many of the collected sentences contained lexical or grammatical errors. To create a parallel corpus of misspelled sentences and their error-free counterparts, we used a Wolof rule-based spell correction tool (Cissé and Sadat, 2023) to generate a file containing the corrected forms of the sentences. We then manually proofread the generated file to correct any remaining grammatical and lexical errors.
For sentences that were initially error-free, we introduced various typographical errors. Most of the introduced errors involve duplication, omission, transposition, or substitution of characters. Table 2 provides an overview on the typos introduced.

Once our synthetic parallel corpus was completed, we were faced with a crucial decision before embarking on the data preprocessing and model training phase, as we needed to determine the

[2] https://www.wolof-online.com

[3] https://twitter.com/SaabalN

[4] https://www.facebook.com/wolofakxamle

[5] http://biblewolof.com

Initial word	Typo category	Misspelled word
Waxtu	Duplication	Waxxtu
Bunt	Omission	Bnt
Juddu	Transposition	udJdu
Ñëw	Substitution	Gneuw
Xaar	Substitution + Omission	Khare
Jàppale	Substitution + Omission	Diapalé
Caabi	Substitution + Omission	Thiabi
Sàkk	Substitution	Spkk

Table 2: Types of errors

atomic linguistic unit that the model will operate on. A substantial number of NLP models have traditionally used tokens as their smallest unit. However, an emerging trend has been noted towards the use of subword units (Sennrich et al., 2016b) as the fundamental building blocks.

The notion of using words as inputs to our model initially appears to be a logical default strategy, mirroring the approach observed in numerous NLP models. However, when applied to spell correction, the token approach can become overly complicated, owing to potential inaccuracies stemming from punctuation use. Additionally, the necessity for NLP models to function on a fixed vocabulary implies that our spell correction tool's vocabulary would need to be comprehensive enough to include every single possible misspelling of every single word encountered during the training process. The implications of this requirement would result in a costly model, both in terms of training and maintenance.

In consideration of these factors, we have decided to use the character as the fundamental building block for our spell checker. This approach has proven to be very effective in translation tasks by Lee et al. (2017). The adoption of character-level segmentation also allows us to preserve a manageable vocabulary size.

For experimental purposes, the overall dataset is divided into three subsets: a training set, a validation set and a test set. We randomly selected 10% of the generated corpus to form the validation and test sets. This was done to make sure that these sets accurately represent the entire dataset. The leftover 90% of the data was then used to create our training set.

3.2. Model architecture

In this study, we employed a customized Transformer model architecture (Vaswani et al., 2017) for the task of Wolof spell correction. The Transformer model has demonstrated remarkable success in various natural language processing tasks by leveraging self-attention mechanisms, which allow it to efficiently process input sequences without the need for recurrent or convolutional layers.

Our model consists of two components: an encoder and a decoder, each comprising five identical layers (Biljon et al., 2020). The encoder's primary task is to manage the input sequences containing misspellings, while the decoder focuses on producing output sequences without misspellings, as illustrated in Figure 1.

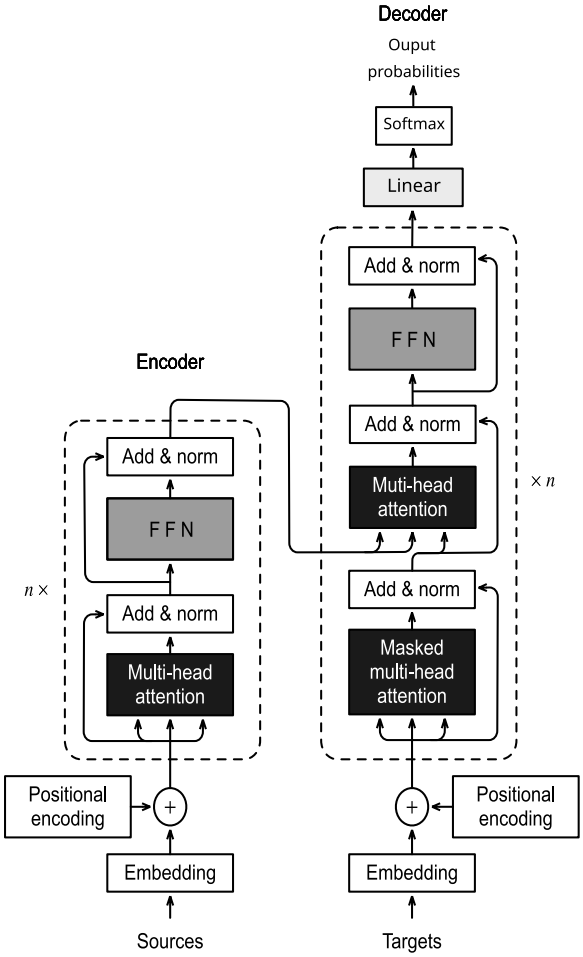


Figure 1: Transformer model(Vaswani et al., 2017)

During the encoding phase, each input word is converted into a vector representation using an embedding layer. To incorporate positional information into the input embeddings, positional encoding is applied. In both the encoder and decoder components of the model, each layer comprises a multi-head self-attention mechanism with two attention heads. This is followed by position-wise feed-forward networks (FFNs) with a hidden size of 256 and a feed-forward size of 1024.

The self-attention process involves generating query (**Q**), key (**K**), and value (**V**) vectors from the input. These vectors are then used to compute a score matrix by performing matrix multiplication between the query and the key vector. The resulting matrix is scaled by the square root of the key vector dimension (d_k). To obtain attention weights, the score matrix is normalized using softmax, representing the importance assigned to different parts

of the input sequence. These attention weights are utilized to derive an output vector, as demonstrated in Eq. 1 (Vaswani et al., 2017). To enable efficient training and stable gradients, residual connections and layer normalization are implemented throughout the network.

$$Attention(Q, K, V) = Softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V \quad (1)$$

The decoder includes two multi-headed attention blocks within a single layer: one for the target sequences and another for the encoder's output. The former multi-head attention is masked to prevent computing attention scores for subsequent words. The latter multi-head attention layer employs the encoder's outputs as queries and keys, while the outputs of the first multi-headed attention layer serve as values. This mechanism empowers the decoder to determine the encoder inputs that are most relevant to its generation process, thereby producing an output sequence without any misspellings. The output from the final pointwise feed-forward layer is then forwarded to a linear layer, serving as a classifier, followed by a softmax layer to generate the corrected text.

For initialization, we employ Xavier initialization with a gain of 1.0 (Glorot and Bengio, 2010) for all trainable weights, while the bias terms are initialized with zeros. The embeddings undergo Xavier initialization with a distinct gain of 1.0. To minimize the number of trainable parameters, a common practice is to tie the source and target embeddings, as well as the softmax layer (Press and Wolf, 2017). Since our model operates at the character level, the default vocabulary size is relatively small. We set the embedding dimension to 256 in both the encoder and decoder, which corresponds to the hidden size of the Feed-Forward Network (FFN) for compatibility. Furthermore, we scale the embeddings by the square root of their size.

To address the issue of overfitting, we employ dropout techniques on various Transformer components. Initially, we apply an embedding dropout rate of 10^{-1} to the encoder and decoder, which helps in dropping words from the embedding matrix (Gal and Ghahramani, 2016). Furthermore, we apply dropout only within the decoder layers at a rate of 3×10^{-1} (Srivastava et al., 2014).

3.3. Model Training

The model training procedure was carefully designed, considering various parameters to ensure rigorous and repeatable results.

We employed deterministic training by using a fixed random seed of 42. To optimize the model, we chose the widely used Adam optimizer (Kingma and Ba, 2015), which features adaptive learning rates and momentum-based parameter updates. For the first and second-order moments, we assigned beta values of $[9 \times 10^{-1}, 999 \times 10^{-3}]$, respectively. The learning rate was initialized at 10^{-4}, and a minimum threshold of 10^{-8} was set to terminate training upon convergence or near-convergence.

To optimize the learning process, we adopted a plateau-based scheduling strategy (Smith, 2017). With a patience value of 5, the learning rate was reduced by a factor of 7×10^{-1} if the validation score did not improve over five consecutive validation rounds. This dynamic adaptation of the learning rate, based on performance feedback, led to enhanced convergence and optimization.

We facilitated efficient parallel computation during training by using a batch size of 4096 tokens (Ott et al., 2018). The token-based batching approach optimized computational resources by forming batches based on the total number of tokens instead of the number of sentences.

Throughout the entire training process, we placed significant emphasis on the model's ability to generalize and perform well by conducting regular evaluations. To ensure a thorough assessment, we established validation intervals of 2000 updates, covering 50 epochs. We carefully selected this interval, considering that setting a validation frequency that is too high might not provide ample opportunities for the model to learn and improve during validation. Furthermore, excessively frequent validation could lead to extended training times and potentially prematurely terminate the training if the validation patience value is not set high enough. Thus, we decided on the mentioned interval to strike a balance.

Moreover, to enhance our ability to closely track the training process and gain comprehensive insights into the model's development, we implemented a logging frequency of 200 updates.

We implemented early stopping by minimizing our cross-entropy loss function, which is a common approach in model training. Continuously monitoring the loss function allowed us to terminate the training when a new low score was reached, effectively preventing overfitting.

To promote diverse predictions and mitigate overfitting, we incorporated two regularization techniques: label smoothing and weight decay. Specifically, we employed label smoothing with a coefficient of 10^{-1} (Szegedy et al., 2016), and weight decay at a rate of 10^{-4} (Srivastava et al., 2014). Label smoothing is a regularization method that redistributes the probability weight from reference tokens to other vocabulary tokens. By reducing the overemphasis on specific reference tokens, label smoothing fosters diversity in the model's output and helps prevent overconfidence in predictions. Weight decay, also known as L2 regularization, is

a technique used to control the complexity of the model. During training, it reduces the magnitude of the model's weights by adding a penalty term proportional to the weight values to the loss function. This regularization term encourages smaller weight values, preventing the model from overfitting the training data and improving generalization performance.

During training, our primary evaluation metric was the well-established BLEU score (Papineni et al., 2002), which measures the similarity between predicted and reference sequences. For efficient evaluation, we used a token-based batching strategy with a batch size of 1024 tokens.

To manage the length of generated sequences during decoding, we set a maximum output length of 175 tokens. Furthermore, we maintained progress monitoring and validation integrity by consistently printing three validation sentences during each validation run.

4. Evaluations

Evaluating spell-checking and correction systems is a crucial task that will help understand their effectiveness and general applicability. While there is no universally accepted standard for evaluating spellchecking and correction systems, three main methodologies have emerged. These methodologies involve classification metrics, machine translation metrics, and information retrieval metrics.

Classification metrics, such as precision, recall, and F-score, are used to assess the performance of automatic spelling correction systems (Starlander and Popescu-Belis, 2002). Machine Translation metrics, including BLEU score (Papineni et al., 2002), CER or WER (Popović and Ney, 2007), and ChrF++ (Popović, 2015, 2017), are also employed in the evaluation. Additionally, information retrieval metrics like MRR (Mangu et al., 2000) can be used.

Considering that our spell checker operates by translating a source text with errors into its most likely correct form, machine translation metrics are the most suitable for measuring our system's performance. For example, the BLEU metric has been widely used to evaluate spell-checking tools in various studies, including those conducted by researchers like Gerdjikov et al. (2013); Mitankin et al. (2014); Sariev et al. (2014). The WER metric was also used in a study by Evershed and Fitch (2014).

After training and evaluating our model on the test set, our spell checker demonstrated high proficiency in various aspects of spelling correction, as shown in Table 3.

The BLEU score, a measure of how well the corrected text matches the reference text in terms of n-gram overlap, is 83%. This high score indicates that the model is capable of producing text that closely aligns with the reference text in both lexical choice and grammatical structure.

Metrics	Scores
BLEU	0.83
WER	0.08
CER	0.03
ChrF++	0.94

Table 3: Performance measures of the spell checker

The WER of 0.08 signifies that, on average, only 8% of the words in the corrected sentences differ from the reference sentences. Similarly, the CER of 0.03 indicates that the corrected sentences have, on average, only 3% character-level differences from the reference sentences. These metrics highlight the effectiveness of the spell checker in accurately identifying and correcting errors at both the word and character levels.

Furthermore, the ChrF++ score of 94% demonstrates a high level of similarity between the corrected sentences and the reference sentences, considering various factors such as precision, recall, and character-level F-score.

4.1. Error Analysis

In addition to the performance metrics mentioned above, it is crucial to conduct a comprehensive error analysis to gain deeper insights into the behavior of our spell checker. We provide a qualitative evaluation of our model on a selection of misspelled Wolof sentences in Table 4. This table presents corrected sentences alongside their corresponding references.

Predictions	References
Ngir **ya** ma def ántalpareet	Ngir **yaa** ma def ántalpareet
Allemañe dëkk bou mag la	**Almaañ** dëkk bou mag la
Woorlu askan wi ñuy jot ci téere yi	**Wóorlu** askan wi ñuy jot ci téere yi

Table 4: Qualitative evaluation

An examination of errors on a subset of the test data has revealed three primary categories of recurring errors produced by our model.

The first group of errors revolves around the correction of long vowels in words. In the Wolof language, distinguishing between long and short vowels significantly impacts word meanings. However, our model consistently struggles to accurately determine when to substitute a short vowel with a long one, resulting in incorrect corrections.

The second group of errors is related to named entities. Named entities, which often deviate from standard Wolof writing conventions, introduce considerable confusion for the model. In some instances, the model incorrectly assumes that these

named entities are erroneous and attempts to rectify them. In other cases, when specific named entities are indeed misspelled and not part of the vocabulary, the model suggests incorrect corrections.

The third group of errors is associated with accent management. Accents play a crucial role in distinguishing and pronouncing words in Wolof. Our model consistently faces challenges when accurately identifying and reinstating missing accents in words.

These findings underscore the need for further refinement of our spell checker, particularly in addressing the complexities of vowel length, handling named entities, and preserving accents within the Wolof language. Moreover, it is essential to explore potential solutions for mitigating these recurring errors, such as incorporating contextual language comprehension and enhancing the model's ability to discern linguistic nuances.

4.2. Test of significance

To establish the statistical significance of the results derived from our evaluation of the spell checker, we conducted a significance test, comparing our model against the sole existing Wolof spell checker[6] accessible online. The objective of this test is to determine the robustness of the observed performance metrics, ensuring that they are not merely a product of random chance.

Our initial step involved the random selection of 100 Wolof sentences from our constructed corpus. Following this preliminary stage, each chosen sentence was input into both correction systems to observe and analyze the proposed corrections.

Subsequently, the correction proposals generated by both systems underwent evaluation by a native Wolof speaker. The evaluator was kept unaware of the source of each correction to maintain impartiality. The applied grading system was as follows: a score of "3" was assigned to sentences that were perfectly corrected and aligned with the reference sentence. A score of "2" was reserved for corrections that, despite minor errors, preserved the original sentence's intended meaning. Lastly, a score of "1" was given to corrections that were entirely incorrect or inadequate.

In order to summarize the evaluations conducted on all the sentences, we have compiled the results in Table 5, which offers an overview of the distribution of scores attributed to each system.

Given the ordinal nature of the evaluations, we opted for the Wilcoxon signed-rank test as the most appropriate statistical tool to discern whether a statistically significant difference exists between the two systems.

Grade	Existing system	Proposed system
1	36/100 (36%)	0/100 (0%)
2	51/100 (51%)	54/100 (54%)
3	13/100 (13%)	46/100 (46%)

Table 5: Systems grades

For this test, we formulated the following null and alternative hypotheses:

$$\begin{cases} H_0 & \text{: There is no significant difference between the two systems.} \\ H_a & \text{: The neural model is significantly superior.} \end{cases}$$

The Wilcoxon test, initially introduced by Wilcoxon (1945), represents a non-parametric approach widely employed for comparing two paired samples. This method is particularly useful when assumptions regarding data distribution are not met or when dealing with ordinal data. We adopted a standard significance level ($\alpha = 0.05$) for this test, considering a result to be statistically significant if the p-value falls below α. In accordance with this methodology, the results obtained for the W-statistic and the p-value are documented in Table 6.

Metrics	Scores
W-Statistic	0.0
p-Value	4.92×10^{-17}

Table 6: W-Statistic and p-Value

The W-statistic serves as an indicator of the cumulative ranks assigned to differences between paired observations, favoring our neural model. A W-statistic value of 0.0 signifies that, in the majority of the compared instances, our proposed neural system has exhibited superior performance when contrasted with the existing rule-based system.

The p-value reflects the likelihood of obtaining such a pronounced difference between the two systems purely by chance, assuming the null hypothesis to be valid. In the context of our Wilcoxon signed-rank test, the null hypothesis postulates that there is no significant difference in the performance of the two systems. An extremely low p-value, such as the one calculated (4.92×10^{-17}), provides compelling evidence against this null hypothesis (H_0), thereby reinforcing the validity of our alternative hypothesis (H_a).

5. Limitations

Our spell checking system has demonstrated good performance, as indicated by its high BLEU and

[6] https://github.com/TiDev00/Wolof_SpellChecker

ChrF++ scores, as well as the relatively low WER and CER scores. However, there are still limitations that require further investigations and improvements.

Firstly, character-level models, such as the one used in this study, are inherently complex and can be time consuming to train. This is due to the larger sequence of data they need to learn from, compared to word-level models. The computational cost of training such models can be particularly high when working with large datasets or languages with extensive character sets.

Secondly, our model may struggle with capturing long-range dependencies within the text. The dependencies between words in a sentence, which often span across several characters, can be difficult for character-level models to understand. This could potentially affect the model's performance in cases that require a deep understanding of sentence-level semantics.

Thirdly, our model lacks the advantage of leveraging pre-trained word embeddings, which capture semantic and syntactic relationships between words. As a result, the model's semantic understanding may be less nuanced compared to models that operate at the word level.

Fourthly, character-level models can be more sensitive to noise in the input data. Spelling errors, inconsistent punctuation usage, and other forms of noise can have a more significant impact on these models, which could lead to lower performance in certain situations.

Additionally, while our model is designed to handle any language that utilizes an alphabet similar to that of the Wolof language, it may struggle with languages that rely heavily on word order. This is due to the model's lack of word-level understanding, which could help in these situations.

Lastly, our model may face difficulties with disambiguation. For instance, words spelled the same but with different meanings can pose a challenge for character-level models, as these models lack access to word-level semantic information.

Given these considerations, there are several areas that could be targeted for improvement. Firstly, the model could be further trained on a wider variety of textual data in order to improve its capacity to handle of less common or more complex errors. Given our current focus on a language with limited available resources, the use of the back-translation technique emerges as a promising strategy. This approach has consistently demonstrated its effectiveness in various domains, such as Statistical Machine Translation (SMT) (Bojar and Tamchyna, 2011), supervised Neural Machine Translation (Sennrich et al., 2016a), and unsupervised Machine Translation (Lample et al., 2017).
In the context of spell-checking and correction, adopting this approach would involve training a model to intentionally introduce a substantial number of realistic spelling errors within clean text. Subsequently, the resulting corpus of corrupted text can be employed to refine our spell checking model.

Furthermore, we suggest further exploration of hybrid models that combine the benefits of both character-level and word-level processing. Such models could potentially leverage the granularity of character-level models while still maintaining a higher-level understanding of word and sentence semantics.

Lastly, considering the computational expenses associated with character-level models, it would be beneficial to conduct research on more efficient training methods. By doing so, we can mitigate the computational burden and improve the overall efficiency of the training process.

6. Conclusion

The present study represents significant progress in the field of automatic spelling correction, particularly for under-resourced and under-represented spoken languages. Our model, which utilizes a transformer-based architecture has produced encouraging results across several evaluation metrics, including BLEU, WER, CER and ChrF++. These outcomes highlight the potential of advanced deep learning techniques to overcome the challenge of spelling errors, even in languages with limited available data.

Despite these promising results, our work has also highlighted certain areas of improvement that could further refine the performance of the proposed system. Our model, being character-level, exhibits certain limitations such as computational complexity, difficulty in capturing long-range dependencies, and sensitivity to noise in the input data. Moreover, the lack of word-level understanding could lead to potential difficulties with languages that heavily rely on word order or face challenges with disambiguation. Furthermore, the investigation of hybrid models, combining the benefits of both character-level and word-level processing, could be a promising direction for future work.

We hope that our findings will encourage further research in this direction, ultimately contributing to the broader goal of building inclusive and effective natural language processing tools for all languages.

7. Bibliographical References

David Adelani, Jesujoba 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 Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 2022. 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: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Sharefah Al-Ghamdi, Hend Al-Khalifa, and Abdulmalik Al-Salman. 2023. Fine-tuning BERT-Based pre-trained models for arabic dependency parsing. *Applied Sciences*, 13(7).

Susan Armstrong, Graham Russell, Dominique Petitpierre, and Gilbert Robert. 1995. An open architecture for multilingual text processing. In *From Texts to Tags: Issues in Multilingual Language Analysis. Proceedings of the ACL Sigdat Workshop.*, pages 30–34, Dublin.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Elan Van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *CoRR*, abs/2004.04418.

David Boilat. 1858. *Grammaire de La Langue Woloffe*. Imprimerie impériale, Paris.

Ondřej Bojar and Aleš Tamchyna. 2011. Improving translation model by monolingual data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland. Association for Computational Linguistics.

Eric Church. 1981. Le Système Verbal du Wolof. Technical report, Université de Dakar, Dakar.

Mamadou Cissé. 2004. *Dictionnaire Francais-Wolof*, 2.éd. révisée et augmentée edition. Dictionnaires des Langues O. Langues et MOndes, L'Asiatheque, Paris.

Thierno Ibrahima Cissé and Fatiha Sadat. 2023. Automatic spell checker and correction for under-represented spoken languages: Case study on Wolof. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Pathé Diagne. 1971. *Grammaire de Wolof Moderne*. Presence Africaine, Paris.

Pathé Diagne. 1997. *Al Xuraan ci Wolof*. Harmattan ; Sankoré, Paris : [Dakar].

Amadou Dialo. 1981. *Structures Verbales Du Wolof Contemporain*. Centre de Linguistique Appliquée de Dakar, Dakar.

Ibrahima Diouf, Cheikh Tidiane Ndiaye, and Ndèye Binta Dieme. 2017. Dynamique et transmission linguistique au Sénégal au cours des 25 dernières années. *Cahiers québécois de démographie*, 46(2):197–217.

Jean Léopold Diouf and Tōkyō Gaikokugo Daigaku. Ajia Afurika Gengo Bunka Kenkyūjo. 2001. *Dictionnaire wolof : wolof-français, français-wolof*. Institute for the Study of Languages and Cultures

of Asia and Africa (ILCAA), Tokyo University of Foreign Studies, Tokyo.

Jean Léonce Doneux. 1978. Les liens historiques entre les langues du Sénégal. *Réalités africaines et langues française: Bulletin du Centre de la Linguistique Appliquée de Dakar*, 7:6–55.

Melynda B. Dunigan. 1994. *On the Clausal Architecture of Wolof*. Ph.D. thesis, University of North Carolina, Chapel Hill.

David Eberhard, Gary Simons, and Chuck Fennig. 2019. *Ethnologue: Languages of the World*, 22nd edition edition. SIL International, Dallas, Texas.

John Evershed and Kent Fitch. 2014. Correcting Noisy OCR: Context Beats Confusion. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 45–51, New York, NY, USA. Association for Computing Machinery.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 1027–1035, Red Hook, NY, USA. Curran Associates Inc.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. 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)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Elodie Gauthier, Laurent Besacier, Sylvie Voisin, Michael Melese, and Uriel Pascal Elingui. 2016. Collecting Resources in Sub-Saharan African Languages for Automatic Speech Recognition: A Case Study of Wolof. *International Conference on Language Resources and Evaluation*, pages 3863–3867.

Stefan Gerdjikov, Petar Mitankin, and Vladislav Nenchev. 2013. Realization of common statistical methods in computational linguistics with functional automata. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 294–301, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, volume 9 of *JMLR Proceedings*, pages 249–256. JMLR.org.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Daniel Hládek, Matúš Pleva, Ján Staš, and Yuan-Fu Liao. 2019. Sequence to sequence convolutional neural network for automatic spelling correction. In *Proceedings of the 31st Conference on Computational Linguistics and Speech Processing (ROCLING 2019)*, pages 102–111, New Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Omar Ka. 1981. *La Derivation et La Composition En Wolof*, volume 77 of *Les Langues Nationales Au Senegal*. Centre de Linguistique Appliquée de Dakar, Dakar.

Mark D. Kernighan, Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 2*, COLING '90, pages 205–210, USA. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Nitin Madnani, Michael Heilman, Joel Tetreault, and Martin Chodorow. 2012. Identifying high-level organizational elements in argumentative discourse. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 20–28, Montréal, Canada. Association for Computational Linguistics.

Max Mangold. 1977. *Wolof Pronoun Verb Patterns and Paradigms*. Number Bd. 3 in Forschungen Zur Anthropologie Und Religionsgeschichte. Homo et Religio, Saarbrücken.

Lidia Mangu, Eric Brill, and Andreas Stolcke. 2000. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *CoRR*, cs.CL/0010012.

Petar Mitankin, Stefan Gerdjikov, and Stoyan Mihov. 2014. An Approach to Unsupervised Historical Text Normalisation. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, pages 29–34, New York, NY, USA. Association for Computing Machinery.

El Hadji Mamadou Nguer, Mouhamadou Khoule, Mouhamad Ndiankho Thiam, Mbaye Baba Thiam, Ousmane Thiare, Mame-Thierno Cissé, and Mathieu Mangeot. 2015. Dictionnaires wolof en ligne : état de l'art et perspectives. In *Colloque National Sur La Recherche En Informatique et Ses Applications*, Thiès, Senegal.

Codu Mbassy Njie. 1982. *Description Syntaxique Du Wolof de Gambie*. Les Nouvelles Editions Africaines, Dakar.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: Words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Maja Popović and Hermann Ney. 2007. Word error rates: Decomposition over POS classes and applications for error analysis. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Stéphane Robert. 2011. Le wolof. *Bulletin de la Société de Linguistique de Paris*, 81.

J. David Sapir. 1971. West Atlantic: An inventory of the languages, their noun class systems and consonant alternation. In Thomas Albert Sebeok, editor, *Current Trends in Linguistics, 7: Linguistics in Sub-Saharan Africa*, number 7 in Current Trends in Linguistics 7 (Ed. by T. Sebeok), pages 45–112. Mouton & Co., The Hague & Paris.

Andrey Sariev, Vladislav Nenchev, Stefan Gerdjikov, Petar Mitankin, Hristo Ganchev, Stoyan Mihov, and Tinko Tinchev. 2014. Flexible Noisy Text Correction. In *2014 11th IAPR International Workshop on Document Analysis Systems*, pages 31–35, Tours, France. IEEE.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. *CoRR*, abs/1704.04368.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Leslie N. Smith. 2017. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017*, pages 464–472. IEEE Computer Society.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle,

Washington, USA. Association for Computational Linguistics.

Radu Soricut and Franz Och. 2015. Unsupervised morphology induction using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1627–1637, Denver, Colorado. Association for Computational Linguistics.

Alexey Sorokin, A. Baytin, I. Galinskaya, E. Rykunova, and T. Shavrina. 2016. SpellRuEval: The first competition on automatic spelling correction for Russian. In *Proceedings of the Annual International Conference "Dialogue"*, volume 15, Moscow, Russia.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Marianne Starlander and Andrei Popescu-Belis. 2002. Corpus-based evaluation of a French spelling and grammar checker. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

Andreas Stolcke. 2000. Entropy-based pruning of backoff language models. *CoRR*, cs.CL/0006025.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, Cambridge, MA, USA. MIT Press.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA. IEEE.

Oscar Täckström, Ryan McDonald, and Jakob Uszkoreit. 2012. Cross-lingual word clusters for direct transfer of linguistic structure. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 477–487, Montréal, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A. Viterbi. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80.

William André Auquier Wilson. 1989. Atlantic. In John Bendor-Samuel, editor, *The Niger-Congo Languages: A Classification and Description of Africa's Largest Language Family*, pages 81–104. University Press of America, Lanham, MD.

David Yarowsky, Grace Ngai, and Richard Wicentowski. 2001. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, HLT '01, pages 1–8, USA. Association for Computational Linguistics.

Lateral Inversions, Word Form/Order, Unnamed Grammatical Entities and Ambiguities in the Constituency Parsing and Annotation of the Igala Syntax through the English Language

Mahmud Mohammed Momoh
Prince Abubakar Audu University
Anyigba, Kogi state, Nigeria
Mahmoodmohammed19@yahoo.com

Abstract

The aim of this paper is to expose the structural form of the Igala language and the inherent complexity related to the translation of the language to a second language vis-à-vis the English language through a configurational probing of its word order, lateral inversions, and unnamed grammatical entities in relation to parsing and annotation in computing. While this study finds out that there is a preponderance of a linguistic typology with subject-verb-object word order and the near total absence of preposition in the speech composition of the Igala language, this fact has not been taken as a serious subject for intellectual consideration. In this study, the abstruseness or incongruity associated with interpreting the Igala syntax through part-of-speech (POS) tagging in relation to its word order, lateral inversion of some phrases, and unnamed grammatical entities (i.e. preposition) in its speech processing into English shall be exposed. Thus, generating a comprehension model for automotive identification, application and/or conversion of these structural forms to the English language shall be the focus of this paper

Keywords: lateral inversion, word order, unnamed grammatical entities, parsing, annotation, Igala, English

1. Introduction

Past works on translation of the Igala language to a second language have focused on the effectiveness of using the English language combined with Igala in teaching in primary schools (Achor and Akor, 2015), evolvement of a modeled language processor that can accept as input Noun Phrases in English language and translate these to Igala (Ayegba, Osuagwu, and Okechukwu, 2014), example acquisition (alignment), matching and recombination (Joshua, Ayegba, and Ojochegbe, 2020), syntactic interference (Attabor, 2019), and contrastive analysis on the use of conjunction (Abraham, 2017). It is worth noting that, no special focus has been placed on the unnamed grammatical entities, word ordering, and the parsing and annotation of inherent syntactic structures. This is notwithstanding the fact that, variations in grammatical rules, word forms and syntactic sequences could be a source of ambiguity and difficulty in translation and comprehension from Igala vis-à-vis the English language by both the machines and the physical learners. This sort of ambiguity has been proven in a more typical sense in regard to translation from a pro-drop language like Japanese or Korean to a non-pro-drop equivalent like English (Wang, Tu, Zhang, 2017). Although, I found out that despite the fact that the Igala language like the English language (see Dryer),[1] French (Bonami, Godard, and Marandin, 1999), Italian (Brunato and Dell-Orletta, 2017), (Namboodiripad, Kim, and Kim, 2017), anchors mainly on a single word order (i.e. subject-verb-object (SVO)), there was still translational ambiguity in implementing an accurate syntactic parsing and annotation for the two languages. Ambiguity in translation from Igala to the English language aside, this sort of mismatch in parsing and annotation could be more serious when carrying out machine-based translation (MT) between Igala and the other languages with contrastive or differential word order such as Korean (Minhui, and Emily, 2015) which uses the postpositional speech form (Mun and Desagulier, 2022) or as in Afaan Oromo (Meshesha, and Solomon, 2018), verb-object-subject (VOS) order as in Malagasy (Ileana, and Postdam, 2024), verb-subject-object (VSO) order as in Welsh (Borsley, Tallerman, and Willis, 2007) or Old Irish (McCone, 1997), or object-verb-subject (OVS) word order as with the not so popular Cariban language; Hixkaryana (Kalin, 2014), in Brazil.

Unlike the observation by Minhui, and Emily (2015) and Namboodiripad, Kim, and Kim (2017) for the Korean language as well as another observation by Fransen (2020) for Old Irish concerning the inherence of multiple word ordering format, I found that the Igala language dwells mainly on a single word order, i.e. Subject-verb-object, as in the phrase; *ū l'ōpā* ≡ "I chewed groundnut" which has the same grammatical approximation in meaning and word sequence with English. However, a notable challenge bedeviling the parsing and annotation of the Igala syntax, most especially with its conversion to English is that of the lateral inversion of some syntactic forms and phrases as

[1] https://www.acsu.buffalo.edu/~dryer/DryerWalsSOVNoMap.pdf

The Fifth Workshop on Resources for African Indigenous Languages @LREC-COLING-2024 (RAIL), pages 152–162

I did observe in respect to this; "*ókwō wē wā*", which is sequentially or literally; "grandparent your came", but actually; "your grandparent came" in the English language.

More also, despite the translation complexity that arises from translation and language teaching when a given part of speech existing in one language does not exist in a corresponding language, from my findings, there are no clearly defined prepositions (which together with postpositions was sometimes referred to as the non-lexical heads of phrases) (Frazier, 1980) in Igala, and thus resulting in incomplete sequential word outlays, vagueness or obscuration of the basic order typology of natural languages and unclear understanding due to this lack of word alternatives during parsing, annotation and general translation as Boquist (2009)[2] did also observed. In this paper, Igala syntactic forms lacking or not containing prepositions would be parsed through parse trees and the corresponding annotations would be converted to the English language as a way of exposing gaps in correspondence and determining the accuracy of translation.

Following the successes of Warren Weaver in the 1950s and the successes that have been recorded in machine translation thereafter – especially in the aspect of part-of-speech tagging in machine translation as Guidivada, and Arbabifard (2018) did rightly observed, I was able to parse and annotate the syntactic structure of the Igala through the English language. Acting upon the suggestion of Guidivada, and Arbabifard (2018) and Jurafsky and Martins (2009), a *transfer-based approach* which uses a three step process was adopted in the segmental structuring of this paper. First, some syntactic analysis (e.g., building a parse tree) is performed on the source text. Second, the syntactic structure is converted (i.e. transferred) into a corresponding structure in the target language. Finally, output is generated from the syntactic structure of the target language. The orthographic frame used in this work as well as the rule of elision expressed in subsection 2.2 conforms to the form adopted in Momoh (2023) and the video on.[3] Furthermore, the triple bar symbol was used to represent equivalence in translation from English to Igala while the approximately equal to symbol ≅ was used to express syntactic isomorphism in translation of syntactic form having differences in lateral sequence of words between Igala and English but the same meaning upon translation to the English language. Using segmented Treebank, six trees-bearing graphs were designed where figures I, II and III contains expressions in the Igala language while IV, V, and VI deals with the English language. Thereafter, the output of the parsing and annotation done was used in demonstrating the structural form of the Machine-based translator being proposed in this paper.

2. Syntactic Analysis, Parsing and Annotation in Igala

In a holistic sense, the Igala language mainly uses the subject-verb-object word order.

2.1 Syntax Parsing and Annotation of the Igala Inverse Possessive Determiners using English

Before designing a parse tree to demonstrate this form of word sequencing in the Igala syntax, the three pronouns; *mā* (their), *mī* (my), and, *nwū* (his/her/it) are considered in respect of their syntactic applications to the Igala phrases demonstrated in the three forms.

Ōmā mā kwū ōrōkā ōnālē ≡ "child their died afternoon yesterday" ≅ "their child died yesterday afternoon";

Ĭyē mī wā ≡ "mother my came" ≅ "my mother came"; and

ẹ́wó nwū dē ≡ "goat his/her be goat" ≅ "this is his/her".

I then did the parsing using the first of the three possessives (i.e. *mā*). The first sentence – '*Ōmā mā kwū ōrōkā ōnālē*' was represented by the parse tree in Figure I;

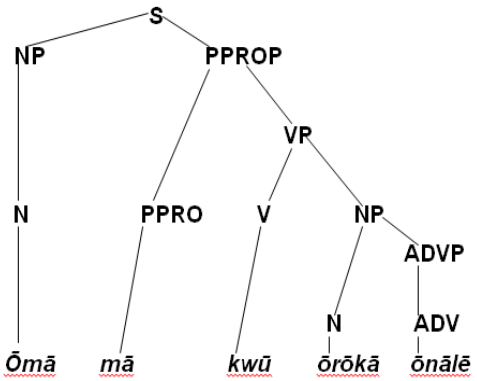


Figure I: Parsing of inverse syntactic determers in Igala.

The next step which was in line with the model proposed in[4] (a system used by the Penn

[2] https://digitalcommons.liberty.edu/cgi/viewcontent.cgi?article=1106&context=honors
[3] https://doi.org/10.48448/e0np-e385
[4] https://www.lancaster.ac.uk/fss/courses/ling/corpus/Corpus2/2PARSE.HTM#:~:text=This%20term%20alludes%20to%20the,article%2C%20P%3Dpreposition.)

Treebank project) (Marcus, Kim, and Marcinkiewicz et al 1994; Santorini, 1990) was to provide a bracket-based morphsyntactic annotation using underscore character (_) in the form of part of speech tags and the use of square brackets annotated at the beginning and the end with the phrase type [s……] as thus:

[S [NP *Ōmā*_NP1 NP] [PPROP *mā*_ PPROP [VP *kwū*_VVD [NP *ōrōkā*_NN1 NP] ADVP_ *ōnālē*] S]

This was also written alternatively as;

[S
[NP *Ōmā* NP]
[PPROP *mā*
[VP *kwū*
[NP *ōrōkā* NP]
[ADVP *ōnālē*]
S]

2.2 Subject-Verb-Object Word Order Parsing and Annotation in Igala

The form of word ordering used in this sub-section follows the same pattern as in the English language. Some phrases and sentences use subjective personal pronouns 'I', 'you' (both in the singular and in the plural form), he/she/it, we, they, and who. While I provided sentences bearing these forms of subjective personal pronouns with respect to these being objects of sentences, both the Treebank and the annotation with respect to this form of word order was done using the subjective form of proper nouns and common nouns by which I provided only one example.

Subjective personal pronouns as subject of the sentence examples:

ọ́mī k'ōmāgọ́lọ́ (I plucked mango) which is simplified albeit ⃠ (unconventionally prohibited in writings) as *ọ́mī kā ōmāgọ́lọ́* (*ọ́mī* + *kā* + *ōmāgọ́lọ́*) = (me + plugged + mango);

ē/me wė ālū (you (singular)/you (plural) + shut + mouth), translated literarily as (*ē/me* + *wė* + *ālū*) = (you/you + shut + mouth);

ī w'ūnyī (she/he/it came home) which is simplified albeit ⃠ as *ī wā ūnyī* (*ī* + *wā* + *ūnyī*) = (she/he/it + came + home);

āwā d'ūnyi (we be home/we are home) which is simplified albeit ⃠ as *āwā* + *dē* + *ūnyi* (we + be + home);

āmā d'ōbē (they took the knife) which is simplified albeit ⃠ as *āmā* + *dū* + *ōbē* (they + took + knife); and,

ēnē k'āfē? (Who took the cloth?) Which is simplified albeit ⃠ as *ēnē* + *kọ́* + *āfē?* (Who + took + cloth?).

The next step that I took was to frame a sentence with a proper noun as the subject of the sentence and a common noun as the sobject of the sentence as was done in https://www.lancaster.ac.uk/fss/courses/. This was done because the form of word ordering considered in this subsection follows the same word order as English which was the language annotated in https://www.lancaster.ac.uk/fss/courses/. The sentence '*Ūgbėdė gw'ọ́jī ọ́dė kā*' which translates as (Ugbede sat on a stool) is the example used. I found out that should the so called prevailing rule on 'conventionality' which adopts apostrophe (as in the word *gw'ọ́jī* above) to fuse two words to one should win through or remain consolidated with respect to machine translation (MT), morphosyntactic annotation of texts becomes complicated. As in the case of the so called 'phrase' *gw'ọ́jī* which can be split to the two separate words gwu which means 'sit' in English and *ọ́jī* which also means 'head' in English but also used to mean 'on', 'above' or 'over' in respect to the dual fusion '*l'ọ́jī*' (pass head) in a more figurative sense (or 'went over' in an actual sense) because of a want of alternative word for expressing the word 'on'.

Thus, in following with the call for the "expansion of contracted forms of multiple words, so that all the words have well defined grammatical categories",[5] in annotating the sentence, 'Ugbede sat on a stool', I used the so called 'unconventional' form of writing the sentence '*Ūgbėdė gwū ọ́jī ọ́dė kā*' rather than '*Ūgbėdė gw'ọ́jī ọ́dė*'. The reason being that the former (i.e. *Ūgbėdė gwū ọ́jī ọ́dė kā*') is amenable to parsing and annotation as it is in line with the Penn Treebank Project while the latter (i.e. '*Ūgbėdė gw'ọ́jī ọ́dė kā*') is not. Here too, the indefinite article 'a' was substituted with the indefinite pronoun '*kā*' which translates in English as 'one'. Although, I found out that articles are classified as separate part of speech in their own right but since they are also considered as a kind of determiners and the word 'one' can be used as a determiner, reference to the word 'one' as used in the sentence is classified as an article and treated as such in the Treebank presented in Figure II.

[5] https://www.cs.rochester.edu/u/brown/242/assts/termprojs/micha/docs/parser.html

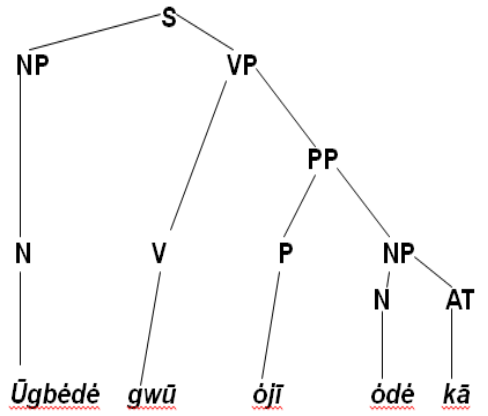


Figure II: Subject-verb-object word order parse tree in Igala.

Next, a second bracket-based morphsyntactic annotation using underscore character (_) in the form of part of speech tags and the use of square brackets annotated at the beginning and the end with the phrase type [s……] was provided in respect of Figure II as thus;

[S [NP *Ūgbėdė*_NP1 NP] [VP *gwū*_VVD [NP *ọ́jī*_II [NP *ọ́dė*_NN1 *kā*_AT1 NP] NP] VP] S]

This was also written alternatively as;

```
[S
    [NP Ūgbėdė NP]
     [VP gwū
          [NP ọ́jī
               [NP ọ́dė kā NP]
            NP]
      VP]
S]
```

2.3 Unnamed Prepositional Entities and Constituency Parsing and Annotation in Igala

In this subsection, I made reference to the 9th Edition of the *Oxford Advanced Learner's Dictionary of English* in which the word preposition was defined as - "a word or group of words such as *in, from, to, out of*, and *on behalf of*, used before a noun or pronoun to show place, position, time or method" (Hornby, 2015). Added to this five (in, from, to, out of, and on behalf of) examples of preposition above were eleven more examples culled from,[6] that included; "beneath," "beside", "between", "in front of", "inside", "near", "off", "through", "toward", "under", and "within". Although, there is the argument adduced by Ilori (2015) to support his claim that there are named prepositions as part-of-speech in Igala for which he went as far as counteracting the claims adduced by other writers like Atadoga (2011) and Ikani (2011) regarding the use of body parts as prepositions but through a careful assessment of the prepositional forms in English pointed out from Hornby (2015) above, I found out that, in truth, the syntactic form of the Igala language does not contain preposition in a more specified sense of the word. In this subsection, I shall cite one example drawn from Ilori (2015)'s abstract where he regarded the word '*tū*' as contained in the phrase '*tú unyí un*' in which he probably meant to say that the word '*tū*' specifically implies the English word 'to' when in reality '*tū*' meant 'unpack' or 'unfasten' while '*tū*' in respect to the preposition has no syntactic base and only exists when its 'root' (the 't') is tied with the word '*ūnyī*' (house or home in English) as in the form '*tūnyī*' as I did pointed out in subsection 2.2 with respect to the word(s) '*gw'ọ́jī*' or '*gwū ọ́jī*' and how this form of dual-word contraction through elision or as a matter of convenience could be a source of ambiguity or encumbrance to word encoding in the design of parse trees and annotation.

In the next lines, I shall try to demonstrate how prepositions are unnamed entities in the syntactic framing of sentences in Igala using the five examples of prepositions offered by (Hornby, 2015) above.

With respect to 'in'; "*ọ́dūdū à wa*" (morning + we + come) which actually translates as "in the morning we shall come" or "we come in the morning";

In respect to 'from'; "*ōmō ī kwọ́*' (there + he/she/it + left) which actually means ("he/she/it came from there" in English). In a sense, the verb 'left' is used instead of 'from' in Igala grammar;

Reference has already been made to the word 'to' above so there is no point adding extra expression to that here;

With respect to 'out of'; "*ėfū mā ī kwọ́*" (belly + them + it + came) which actually meant ("out of them it came") in English;

With respect to 'on behalf'; "*t'ọ́dū mī*" (t + name + me) which actually means ("because of me" in English).

Thus, using the first sentence in respect to the word 'in', a simple parse tree with its annotation was provided to shed more light on this. Figure III has the parse tree.

[6] https://academicguides.waldenu.edu/writingcenter/grammar/prepositions#:~:text=%22beneath%2C%22%20%22beside%2C,street%20from%20the%20grocery%20store.

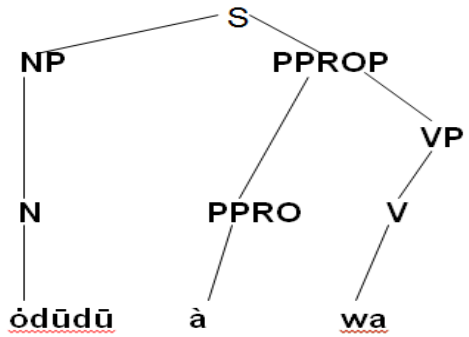


Figure III: Parse tree expressing unnamed preposition in Igala.

I then framed an annotation for the syntactic form of the parse tree in Figure III as thus;

[S
[NP *ódūdū* NP]
[PPRO *à*
[VP *wa*
NP]
VP]
S]

3. Conversion of the Igala Syntactic Form to English

In this section, I converted the parse tree/annotation in the preceding section into English in the form of a translation.

3.1 Conversion of the Igala Possessive Determiners to English

I found that, owing to the lateral inversion of syntaxes, there is a disproportionate incongruity in converting syntactic forms in Igala to English in a figurative sense as demonstrated below. What I did was to reverse the phrases used in the second section above from English to Igala.

Thus, the three forms of pronouns; their (*mā*), my (*mī*), and, his/her (*nwū*) are considered in respect of their syntactic applications in the English phrases given in the example below.

"Child their died afternoon yesterday" ≡ *Ōmā mā kwū ōrōkā ōnālē.*

I then created a parse tree representing this word order in English as thus;

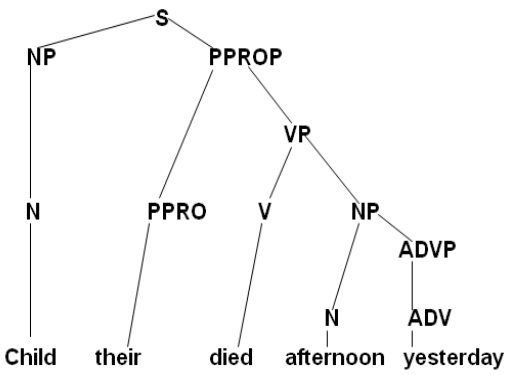


Figure IV: Parse tree showing conversion of possessive determiners from Igala to English.

Figure IV was annotated a thus;

[S [NP Child_NP1 NP] [PPROP their_ PPROP [VP died_VVD [NP afternoon_NN1 NP] ADVP_ yesterday] S]

This was also written alternatively as;

[S
[NP Child NP]
[PPROP their
[VP died
[NP afternoon NP]
[ADVP yesterday]
S]

3.2 Conversion of Subject-Verb-Object Word Order from Igala to English

Notwithstanding the fact that the English subject-verb-object word order also exists in Igala, getting an accurate translation for English to Igala proved a little bit problematic as shown in Figure V.

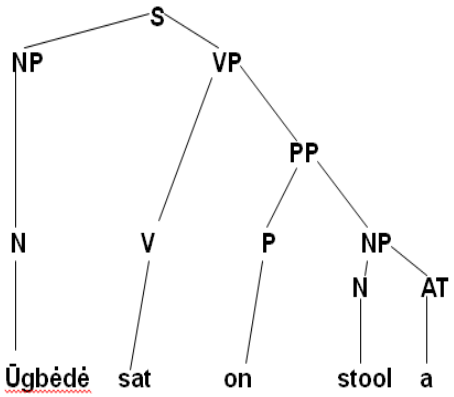


Figure V: Parse tree showing the conversion of subject-verb-object word order from Igala to English.

A bracket-based syntactic annotation for Figure V was given below;

[S [NP Ugbede_NP1 NP] [VP sat_VVD [PP on_ II [NP stool_NN1 a_AT1 NP] PP] VP] S]

This was also written alternatively as;

```
[S
    [NP Ugbede NP]
     [VP sat
          [PP on
               [NP stool a NP]
           PP]
      VP]
S]
```

3.3 Conversion of Igala Syntactic Forms with Unnamed Prepositional Entities through Parsing and Annotation in English

Here, what I did was to copy the phrase used for the parsing and annotation of the sentence bearing the unnamed preposition in subsection 2.3 (i.e. "*ódūdū à wa*" - (morning + we + come) which was used to build a corresponding parsing and annotation in English. This was represented in the constituency parsing on Figure VI and the annotation that comes below it.

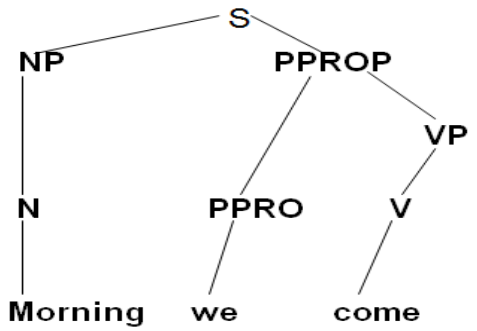


Figure VI: Parse tree showing the conversion of unnamed prepositional entities from Igala to English.

The syntactic form of the phrase on Figure VI was annotated as thus;

```
[S
    [NP ódūdū NP]
     [PPRO à
          [VP wa
            NP]
      VP]
S]
```

4. Output Generated from the Grammatical Structure of the Igala Language

From findings in this work, it becomes clear that annotation of the syntactic form of the Igala language would remain a herculean task. The fact that certain words are being conjoined arbitrarily as one through the use of apostrophe makes it hard for words to maintain their original form during sentence composition, making it hard for their annotation and translation to a second language. The implication of this is the presence of mixed signal while trying to convert syntactic form from the Igala language to English in a more specific sense.

Thus, in using the apostrophe for conjoining two words as one which is currently the case among most writers of the Igala language in which case, the phrase 'leave there' becomes '*kw'ōmō*' /kwomo/ rather than '*kwó ōmō*' /kweu omo/ and 'put there' becomes, '*ṭ'ōmō*' rather than '*ṭō ōmō*' - a practice done as a way of endearing fluency in conversation (Momoh 2023). Being an isolating language agglutinating inflectional morphemes with more than one unit of meanings denoted by separate part-of-speech, how to encode the specific word and them along their individual grammatical unit during parsing and annotation for a working machine-based translation becomes difficult. While most words are classed as having a 1:1 morpheme per word ratio, others like '*k'ōmō*' /komo/ (hit there), '*g'ōmō*' /gomo/ (look there), have a 2:1 morpheme per word ratio that is similar to the explanation provided in respect to Russian by Comer (2021). Following from this fact, this writer found that the syntactic codes for part-of-speech (POS) parsing and annotation proposed in the Penn Treebank Project are insufficient for the parsing and annotation of the Igala language. Whether to use special identifiers such as the plus sign (+) or the slash sign (/) in expressing agglutination, i.e. to express the parsed form of *gomo* (look there) as VP+ADVP or V+ADV and VP/ADVP or V/ADV on the vertical dashes or whether to have words like '*k'ōmō*' /komo/ (hit there), '*g'ōmō*' /gomo/ (look there) written without the use of the eliciting mark expressed by the application of the apostrophe remains an issue of concern. Although the use of the + (plus sign) as suggested here comes with a different mode of application, but this comes close to the same indicator used for analyzing contraction as PPSM+BEM in the pioneering Brown Corpus (Marcus, Santorini, and Marcinkiewicz, 1993).

With respect to the parse tree and annotation of the inversed possessive determiners, I found just a 20 per cent mean correlation in the sequence of word order in the syntactic translation from Igala to English and vice versa, with the result that, out of the five words used apiece, only the median word '*kwū*' and 'died' maintained consistency in the sequence of word arrangement as shown in the third vertical dashes on the two figures (I and IV) representing the parse trees and also on their individual annotations. There was also an attendant displacement of four (*Ōmā, mā, ōrōkā*, and *ōnālē*) of the five Igala words and four (child, their, afternoon, and yesterday) of the five English words upon conversion from Igala to English. Following from this fact, I found an 80 percentage point to this end. More also, owing to the nonexistence of preposition in the Igala word forms, the word '*ójī*' (head) – but could as well be

translated as 'thief' in English and which was represented as a noun on the third vertical dash of the parse tree of Figure II was replaced with the word 'on' – a preposition, upon conversion to English on Figure V. The implication of this is a noun + noun sequence in the syntactic order of the phrase '*ọ́jī ọ́dė́*' as was also done by Ilori (2015) on page 146 of his paper.

It therefore implies that the word '*ọ́jī*' in a more figurative sense would have to be recognized as 'on' during word conversion through parsing and annotation in English and in which case, both the parse tree and the annotation of the phrase '*Ūgbė́dė́ gwọ́jī ọ́dė́ kā*' would have to be redrawn in line with the Brown Corpus format as thus;

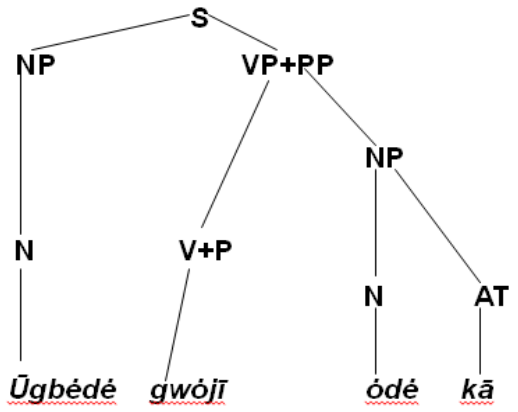


Figure VII: Parse tree showing the splitting of gwọ́jī into gwū and ọ́jī.

[S [NP *Ūgbė́dė́*_NP1 NP] [VP+PP *gwọ́jī*_ II [NP *ọ́dė́*_NN1 *kā*_AT1 NP] VP+PP] S]

This was also written alternatively as;

```
[S
    [NP Ūgbė́dė́ NP]
     [VP+PP gwọ́jī
              [NP ọ́dė́ kā NP]
           NP]
      VP+PP]
S]
```

More also, the Treebank and the syntax annotation with respect to table II and IV shows an 80 per cent correlation in word sequence with the result that, while there is no word elision as it is in respect to figures I and IV, the words 'a' and 'stool' were however inversed laterally from the ordering sequence they exist in the Igala syntactic structure in so that, the syntactic or phraseological form of 'a stool' in the English language became reversed as 'stool a' or 'stool one' ⇔ *ọ́dė́ kā*. Rising from this fact that none of the five indicators represented in figures I through VI was unnamed in the two languages, a 60 per cent translation accuracy using the subject-verb-object for Igala and English was arrived at. I found that from the five syntactic variables exemplified by the five vertical dashes on the parse trees on figures IV and V there was an accuracy in word frequency of 3 > 2 and an error of 2 < 3. The implication of this is that there was a 40 per cent error towards this end.

Through figures III and VI we also noticed two kinds of errors or inaccuracy in translation from the sequence of word on the figures as was also apparent in the flow pattern of the annotation of the content of the parse trees on both figures. There were cases of lateral inversion in translation from Igala to English.

Figure VIII is a two-way crawling translator that can also be a word-to-meaning finder through the pipes connecting A1 and A2 and B1 and B2. Input is received via either side of the translator with the blue colour representing channels for the flow and transmission of words in Igala while the orange colour boxes represent the English equivalent. The vertical rectangle in either portion of the three boxes coded I (input) is the transformer which is connected to eight word banks representing the eight parts of speech in which the corpuses would be fed. A1 and B1 are word receivers while A2 and B2 are parsers but can also function alternatively as input and output processor if the machine is commanded to find word and meaning in the given language.

A1 and B1 are machine-based parsers and/or annotators that decode questions transmitted from the affected I boxes of the source language(s) while A2 and B2 are parse and/or annotation converter into the target language. With the syntax moved through the various I boxes, these are sent into the transformer.

From the eight (8) boxes fed lemmas or words according to the given part of speech category of each lemma, i.e. boxes attached to C1 and D1, these lemmas and their meanings as stored in each of the eight boxes are connected directly through the eight pipes linking the eight boxes to the word receivers attached to the individual transformer.

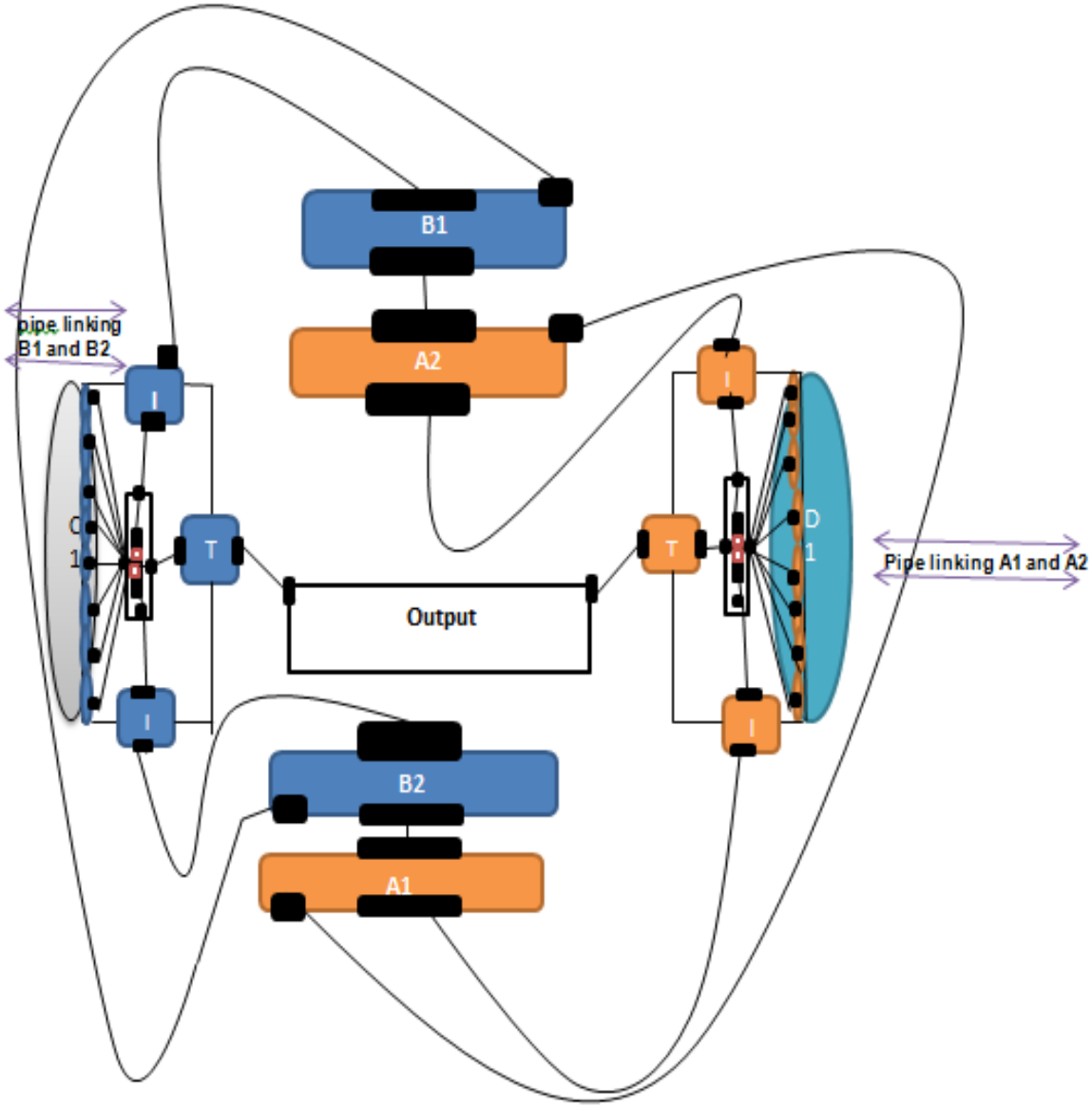


Figure VIII: Bidirectional model for the design of a Machine Translator of English to Igala and from Igala to English.

When syntax received by A2 and B2 are parsed/annotated through the converter to be built into A2 and B2, they are moved through the next I channel on the pipeline to the transformers on either side of the model. Rather than translating the syntax directly, the transformer, using scanners connected to it, finds words equivalents in the part-of-speech boxes, using the scanners and word receivers attached to the individual transformer. Through the word receiver, the individual word in the phrase/syntax sequence are moved into the scanner and then sent into the transformer for scrutiny. To deal with cases of ambiguity, lateral inversion, and unnamed grammatical entities, the transformer shall be trained through part-of-speech tagging, in which case, while recognizing several meanings of a given lemma as shall be drawn directly from C1 and D1. Contextual applications such as *l'ọ́jī* or *lọ́jī* (passed head) in a more literary sense, but actually 'went over', or *gw'ọ́jī* or gwọ́*jī* (sit head) in a more literary sense, but actually 'sit on', upon conversion to English would then be represented in the transformer as *l'ọ́jī* or *lọ́jī* => 'went over', while the syntactic form *gw'ọ́jī* or gwọ́*jī* => 'sit on'. In so doing, the transformer, even though it would receive a wide multiple meanings on the words being fed into it would be able to make prediction on the actual context the translation should appear, so that rather than interpreting a phrase like gwọ́*jī* ẹ̀bī*j*ẹ́ as 'sit head iron', it becomes gwọ́*jī* => {sit on} + {iron} => {sit on iron} => {sit on the iron}. You will notice the inking of the definite article 'the' with a tan tinted background. The reason is that upon parsing/annotation from Igala to English at A2 on figure VIII, the parser/annotator could not identify the word 'the', but since the syntactic form 'sit on chair' did not make a perfect sense in English, the word was

given a separate colour as a way of expressing explicit insertion that are not intrinsic in the word sequence of the source language, upon translation.

In probing for the unnamed grammatical entities, a similar annotation format - the BIO which is used for probing named entities in word prediction is used, although with a different purpose. The BIO annotation format – (inside, outside, beginning) was used in a parallel fashion to detect words and their meanings in the two language. Through the provision of the BIO annotated forms of syntax in the transformer, the transformer would be able to identify word sequence that comes closer to the meaning of another word sequence in the corresponding language, i.e. in respect to the phrase; *kā kū gbō* (say + that I + hear) 'say it let me hear', this phrase would be annotated using the BIO format as; say (B), it (O), let (O), me (O), hear (I). Through the application of a scanner with the capacity to detect both known and unknown entities during translation, it becomes possible for the translating machine to sense words that are not in the word sequence as translated which it thus, marks out using the tan colour pointed out above.

The two scanners attached to the transformers (the two tan-colour boxes attached to the transformer (the box with the vertical rectangle shape tied to C1, the two blue I boxes and the blue T to the left, and D1, the two orange colour I boxes and the yellow T box to the right)), are word detectors. Depending on the application though, either of these tan boxes send words in sequence as received from the transformer from either A2 or B2 through any of the two I on either side of figure VIII, following parsing and annotation. It then sense these words these words and their meanings through a crawling mechanism in more of a sense as the Google Search engine from either of the eight part-of-speech boxes on the two far flanks of the model for tagging and processing into meaning. The second tan boxes inside of the transformer; the two at the top, collects and returns unselected words and meanings from the transformer back to the part-of-speech boxes they evolved from in C1 and D1. When words are processed in the transformer, the translated equivalent are send via a pipe to the two T boxes on either side, for onward transmission to the screen as output.

5. Conclusion

The purpose of this paper therefore sprang from the need to create parsing trees and syntax annotation that could serve as bedrock of input materials that could be used for the development of a language corpus for the Igala, a needful resource that does not 'really' exist because previous efforts by Ayegba et al. (2017) are inadequate for want of extensive modeling required while the paper by Joshua et al., (2020) does contain corpuses built on program interfaces, they are however not centered exclusively to corpuses and so they are not so comprehensive enough to serve the essence of that subject – corpus.

6. Acknowledgement

This was conducted through the encouragement and the technical support that I received from Professor Menno van Zaanen of the Northwest University in South Africa and the rest of his team from the South African Center for Digital Language (SADiLaR). More also, my gratitude goes to the management of the Department of History and Security Studies of the Umaru Musa Yar'adua University (UMYU) in Katsina where am currently a graduate student and the management of the Prince Abubakar Audu University (PAAU), Anyigba, Kogi state, Nigeria, where am currently working for their patience and encouragement through the period of this research – with a special word of 'thank you' for the Vice Chancellor of PAAU, Professor Marietu Ohunene Tenuche and the current Head of Department of History and Security Studies at UMYU, Dr Waisu Safana. It is also worth mentioning the effort of my dear uncle, the late Alhaji Ibrahim Sule who was the hand the held me through school after I lost my dad at just six years of age. The collective effort of the aforementioned persons/groups, either in an immediate or remote sense towards this paper would always remain fresh in my heart.

7. Ethical Statement

While conceiving a work of this nature, there are lots of issues to be dealt with such as the merit from which supportive datasets are extracted, the question as to whether available inputs that can be used to model a machine-based translator such as the tokens, lemmas, and the corpuses for an under-resourced language like the Igala where most of what exists as data needs extra verification before incorporation into other resource forms remains a biting question. Although while the accuracy of presentation and the approach adopted toward modeling by this author is not in doubt in a more literary sense, the fact that the was no provision for parsing contractive or agglutinative word forms by the Penn Treebank P.O.S tags led to the blending of these tags with a separate approach that was evolved by this very author – one that comes close to the tag sets and approach used by the Brown Corpus. Thus, in using the content of this paper, extra care should be taken by the consumers because of certain steps and approaches that might not be so accurate.

8. References

Abraham Sunday Unubi. (2017). "The use of

conjunctions in English and Igala: A contrastive analysis". *International Journal of Advanced Multidisciplinary Research*, 4(8), Pages 34-64.

Atadoga, F. T. (2011). 'Igala Morphology'. In Omachonu G. S (ed.) *Igala language Studies*. Saarbrucken: LAP Lambert Academic Publishing, Pages 76-102.

Attabor Theophilus Ocheja. (2019). "Syntactic Interference: A Study of Ígálá and English Noun Phrases in Malachai 1:6 and Mathew 2:1". *Journal of Literature, Languages and Linguistics*, 60, Pages 40-46.

Ayegba Sani Felix, Osuagwu O.E, and Njoku Dominic Okechukwu. (2014). "Machine Translation of Noun Phrases from English to Igala using the Rule-Based Approach". *West African Journal of Industrial and Academic Research*, 2(1): 18-28.

Ayegba Sani Felix, Abu Onoja, and Musa Ugbedeojo. (2017). "English – Igala Parallel Corpora for Natural Language Processing Applications". *International Journal of Computer Applications*, 171(9): 1-6.

Bonami Olivier, Godard Danièlle, and Marandin Jean-Marie. (1999). "Constituency and word order in French subject inversion". In Gosse Bouma, Erhard Hinrichs, Geert-Jan M. Kriujff, and Richard Oehrle (eds.) *Constraints and Resources in Natural Language Syntax and Semantics*. CSLI Publications, Stanford, USA, Pages 1-20.

Borsley D. Robert, Tallerman Maggie, and Willis David. (2007). *The Syntax of Welsh*. Cambridge University Press, Cambridge, UK, Pages 1-102.

Brunato Dominique, and Dell-Orletta Felice. (2017). "On the order of words in Italian: a study on genre vs complexity". In *Proceedings of the Fourth International Conference on Dependency Linguistics*, Pisa, Italy, September 18-20, Pages 25-31.

Comer J. William. (2021). "Russian's Most Frequent Words and Implications for Vocabulary Instruction," *Russian Language Journal*, 71(1),

Emmanuel Edoja Achor, and Christiana Akor. (2015). "Exploring the Use of Igala Language in Teaching Statistics to Samples of Selected Primary Six Learners". In *Proceedings of ISTE International Conference on "Towards Effective Teaching and Meaningful Learning in Mathematics, Science and Technology",* University of South Africa, Pages 55-66.

Fransen Theodorus. (2020). "Automatic Morphological Parsing of Old Irish Verbs Using Finite-State Transducers". *Leeds Working Papers*, 1: 15-28.

Frazier Lyn. (1980). "Parsing and Constraints on Word Order". University of Massachusetts Occasional Papers, 6(5): 177-198.

Gudivada N. Venkat, and Arbabifard Kamyar. (2018). "Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications". In Gudivada N. Venkat, and Rao, C.R. *Handbook of Statistic 38*. https://www.sciencedirect.com/topics/computer-science/translation-model#:~:text=The%20language%20model%20can%20be,using%20the%20notion%20of%20alignments.

Hornby S. Albert. (2015). Oxford Advanced Learner's Dictionary of English, (Ninth Edition). Oxford University Press, Oxford, UK, Page 1210.

Ikani, F. E. (2011). 'Sense and Meaning Relations in Igala'. Omachonu G. S (ed.) Igala Language Studies. Saarbrucken: LAP Lambert Academic Publishing, Pages152-176.

Ileana Paul, and Postdam Eric. (2024). "Malagasy Framing Demonstratives and the Syntax of Doubling". *Glossa: A Journal of General Linguistics*, 9: 1-37.

Ilori, Folorunso. (2015). "Prepositions in Igala". *Ihafa: A Journal of African Studies*, 7(1): 139-160.

Joshua Attah, Ayegba Sani Felix, and Orah Richard Ojochegbe. (2020). "Example Based Machine Translation of English to Igala Language". *International Journal of Trend in Research and Development*, 7(1): Pages 32-35.

Jurafsky Daniel, and Martin H. James. (2009). Speech and Language Processing. https://en.wikipedia.org/wiki/Transfer-based_machine_translation#cite_note-slp-1

Kalin Laura. (2014). "The syntax of OVS word order in Hixkaryana". *Natural Language and Linguistic Theory*, 32: 1089-1104.

Mahmud Mohammed Momoh. (2023). "Vowels and the Igala Language Resources". In Proceedings of the Fourth workshop on Resources for African Indigenous Languages, pages 106-114.

Marcus Mitchell, Beatrice Santorini, and Marcinkiewicz Mary Ann. (1993). "Building a Large Annotated Corpus of English: The Penn Treebank". *Association for Computational Linguistics*, 19(2), pages 113-330.

Marcus Mitchell, Kim Grace, and Marcinkiewicz Mary Ann., et al. (1994). "The Penn Treebank: Annotating Predicate Argument Structure". https://aclanthology.org/H94-1020.pdf

McCone K. (1997). *The Early Irish verb. Revised edition with index verborum.* An Sagart, Maynooth, Ireland, Page 17.

Million Meshesha, and Yitayew Solomon. (2018). "English-Afaan Oromo Statistical Machine Translation". *International Journal of Computational Linguistic*, 9(1): 26-31.

Minhui Choi, and Emily Schmidt. (2015). "Postpositions and Word Order Variation in Korean". Linguistic Portfolios, 4(10): 108-115.

Mun Seongmin, and Desagulier Guillaume. (2022). "How do Transformer-Architecture Models Address Polysemy of Korean Adverbial

Postpositions?". In Proceedings of Deep Learning Inside Out (DEELIO 2022): The Third Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Dublin, Ireland, Pages 11-21.

Namboodiripad Savithry, Kim Dayoung, and Kim Gyeongnam. (2017). "English-dominant Korean-speakers show reduced flexibility in constituent order". *Chicago Linguistic Society*, 53:

Philpot, W.T.A. (1935). "Notes on the Igala Language". *Bulletin of the School of Oriental Studies*, 7(4): Pages 897-912.

Santorini Beatrice. (1990). *Part of Speech Tagging Guidelines for the Penn Treebank Project.* https://www.cis.upenn.edu/~bies/manuals/tagguide.pdf

Wang, L., Tu, Z., and Zhang, X., et al. (2017). "A Novel and Robust Approach for Pro-drop Language Translation". *Machine Translation*, 31: 65-87.

Author Index

Assabie, Yaregal, 124
Atsbha, Gebregewergs Mezgebe, 94
Azouaou, Faical, 133

Batista-Navarro, Riza, 115
Berg, Ansu, 55
Boubred, Mohamed, 133
Boucetta, Anfal Yousra, 133

Chennoufi, Sara, 133
Cissé, Thierno Ibrahima, 140
Coffey, Joseph R., 20
Cristia, Alejandrina, 20

Demolin, Didier, 1

Eiselen, Roald, 55

Gaustad, Tanja, 55
Gauthier, Elodie, 10
Gelbukh, Alexander, 107
Ghio, Alain, 1
Gidey, Gebrearegawi Gebremariam, 94
Guellil, Imane, 133
Guissé, Abdoulaye, 10

Houichi, Yousra, 133

Ibrahim, Nuhu, 115

Kalita, Jugal, 107
Karani, Michael, 1
Kolesnikova, Olga, 107

Lawrence, Matt, 115

Marais, Laurette, 77
Meynadier, Yohann, 1
Moape, Tebatso G., 32
Momoh, Mahmud Mohammed, 152
Mulford, Felicity, 115

Ndiaye, Aminata, 10
Ngcungca, Nkazimlo N., 45

Ojo, Sunday Olusegun, 32
Olugbara, Oludayo O., 32

Posthumus, Lionel Clive, 77
Pretorius, Laurette, 77
Pretorius, Rigardt, 55

Rudman, Sharon, 45

Sadat, Fatiha, 140
Sibeko, Johannes, 37, 45, 66
Southwood, Frenette, 86

Taffa, Tilahun Abedissa, 124
Teklehaymanot, Hailay Kidu, 94
Tonja, Atnafu Lambebo, 107

Usbeck, Ricardo, 124

van Zaanen, Menno, 66

White, Michelle J., 86

Yalala, Sefela Londiwe, 86