Preface to the proceedings of RAIL 2022

The Third workshop on Resources for African Indigenous Languages (RAIL) was held (in person) on 30 November 2022 in Potchefstroom. It was organized by the South African Centre for Digital Language Resources (SADiLaR), and was co-located with the 10th Southern African Microlinguistics Workshop, which took place from 1 to 3 December 2022.

The RAIL workshop series aims to bring together researchers who are interested in showcasing their research and thereby boosting the field of African indigenous languages. This provides an overview of the current state-of-the-art and emphasizes availability of African indigenous language resources, including both data and tools. Additionally, it allows for information sharing among researchers interested in African indigenous languages and also starts discussions on improving the quality and availability of the resources. Many African indigenous languages currently have no or very limited resources available and, additionally, they are often structurally quite different from more well-resourced languages, requiring the development and use of specialized techniques. By bringing together researchers from different fields (e.g., (computational) linguistics, sociolinguistics, language technology) to discuss the development of language resources for African indigenous languages, we hope to boost research in this field.

The RAIL workshop is an interdisciplinary platform for researchers working on resources (data collections, tools, etc.) specifically targeted towards African indigenous languages. It aims 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.

At the third edition of the RAIL workshop, we received eighteen high quality submissions, which were all reviewed by three reviewers. The reviewing process was double blind. Finally, fourteen submissions were accepted. Note that the aim was to get as many people together to share their ideas and research results, while attaining a high quality event. This lead to a workshop, which consisted of a full day of presentations. This publication adheres to DHET's 60% rule, authors in the proceedings come from a wide range of institutions. In total, the audience could listen to fourteen presentations. Each presentation consisted of 25 minutes (including time for discussion). In fact, after all presentations many interesting discussions took place.

As organizers, we would very much like to thank the programme committee for their in-depth reviewing of the submissions and for providing useful recommendations to the authors.

- Danie Prinsloo, University of Pretoria, South Africa
- Elsabe Taljard, University of Pretoria, South Africa
- Emmanuel Ngue Um, Université de Yaoundé I, Cameroon
- Emmanuel Sithole, South African Centre for Digital Language Resouces, South Africa
- Febe de Wet, Stellenbosch University, South Africa
- Friedel Wolff, South African Centre for Digital Language Resouces, South Africa
- Gonneke Groenen, North-West University, South Africa
- Iris Hendrickx, Radboud University Nijmegen, The Netherlands
- James Akinola, Chrisland University, Nigeria
- Johannes Sibeko, Nelson Mandela University, South Africa
- Marissa Griesel, UNISA, South Africa
- Roald Eiselen, CTexT, South Africa
- Sibonelo Dlamini, University of KwaZulu-Natal, South Africa
- Sree Ganesh Thottempudi, South African

Centre for Digital Language Resouces, South Africa

- Tanja Gaustad, CTexT, South Africa

RAIL 2022 organizers and proceedings editors

- Jessica Mabaso
- Rooweither Mabuya
- Muzi Matfunjwa
- Mmasibidi Setaka
- Menno van Zaanen

Localising the Mozilla Common Voice platform for South Africa's official languages

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University & School of Electrical, Electronic and Computer Engineering, North-West University
fdw@sun.ac.za

Bukula, Andiswa
South African Centre for Digital Language Resources, North-West University
Andiswa.Bukula@nwu.ac.za

Karsten, Willem
School of Electrical, Electronic and Computer Engineering, North-West University
willemkarsten2308@gmail.com

Puttkammer, Martin
Centre for Text Technology, North-West University
Martin.Puttkammer@nwu.ac.za

Schillack, Erwin
School of Electrical, Electronic and Computer Engineering, North-West University
schillackerwin@gmail.com

Wierenga, Rone'
Virtuele Instituut vir Afrikaans
rone@viva-afrikaans.org

Eiselen, Roald
Centre for Text Technology, North-West University
Roald.Eiselen@nwu.ac.za

Abstract

Despite many attempts to address the situation, South Africa's official languages remain under-resourced in terms of the text and speech data required to implement state-of-the-art language technology. To ensure that *no language is left behind*[1], resource development should remain a priority until a strong digital presence has been established for all indigenous languages. This paper provides an overview of previous projects that were specifically aimed at speech resource development and introduces an ongoing initiative to launch South Africa's languages on the Mozilla Common Voice platform.

Keywords: DHASA, under-resourced languages, speech resources, Mozilla Common Voice

1 Introduction

South Africa's constitution recognizes eleven official languages: Afrikaans (Afr), South African English (Eng), isiNdebele (Nbl), isiXhosa (Xho), isiZulu (Zul), Sepedi (Nso), Sesotho (Sot), Setswana (Tsn), Siswati (Ssw), Tshivenda (Ven), and Xitsonga (Tso). With the exception of English and Afrikaans, all the official languages belong to the South-Eastern Bantu family. IsiNdebele, Siswati, isiXhosa and isiZulu are part of the Nguni group of languages and Sepedi, Sesotho and Setswana are part of the Sotho language group. The languages within each family are closely related, with similar orthographic and morphosyntactic attributes.

Most people in South Africa speak more than one Bantu language and English. As a result, English serves as *lingua franca* and is most frequently used in commerce and law. Some of the country's citizens have access to language and speech technology through English. For the ten remaining languages, much remains to be done to match the level of technology development that has already been achieved for languages like English. In this regard South Africa's indigenous languages are in the same position as the majority of the almost 7 000 languages that are spoken in the world today: usable language and speech technology is not readily available yet (Adda et al. 2019, Joshi et al. 2020).

Despite various projects aimed at addressing this situation, the pace of resource development in South Africa's languages has not kept up with the rate at which technology and the data requirements associated with state-of-the-art techniques have advanced. As a result, many of the latest technology, especially deep learning techniques, cannot be implemented effectively for South Africa's local languages due to

a lack of appropriate data.

This paper describes a recent initiative to launch South Africa's official languages on Mozilla's Common Voice[2] platform. The Common Voice project aims to make speech recognition technology open and accessible by creating open, high quality, publicly available data sets in as many languages as possible. For a language to achieve *launched* status, the Common Voice website needs to be localised and at least 5 000 sentences in the target language have to be collected and be available in the open domain under CCo licensing [3]. Once *launched* status is achieved, the sentences are used as prompts for speech data collection through the Mozilla Common Voice platform. While Mozilla makes the Common Voice platform freely available, they are not involved in localisation and data collection and do not provide any financial support to participants. The presence of a language on the platform is determined by language communities themselves.

2 Background

A number of projects have already contributed to the establishment of basic language resources as well as speech and text technology in South Africa's official languages. Many of these were supported by the South African Government[4].

One of the first attempts to develop technology in the country's indigenous languages, the *African Speech Technology* (AST) project, was funded by the Department of Science and Technology's Innovation Fund (Roux et al. 2004). One of the aims of the project was to prepare South Africa's languages for a digital future. It was also envisioned that language technology would facilitate multilingual information access to South Africa's citizens. Five telephone speech databases in isiXhosa, Sesotho, isiZulu, South African English and Afrikaans were developed during the course of the project. The data was transcribed orthographically as well as phonetically and used to develop a prototype version of a multilingual, telephone-based hotel booking system. A limited domain text-to-speech voice was also built for each of the five languages, allowing the system to provide dynamic although domain limited speech feedback.

Subsequent to the AST project, three *Lwazi*[5] projects were funded by the South African Department of Arts and Culture with the aim of extending the available telephone speech data sets to include all 11 official languages and to increase the impact of speech technologies in South Africa (Barnard et al. 2010, Kuun 2012, Calteaux et al. 2013, Titmus et al. 2016). Toward the latter aim, text-to-speech and speech-to-text systems were developed in all 11 languages and evaluated in applications including a voice-based telephone service for rural veterinarians and a multilingual, telephone-based interactive voice response system for the Department of Basic Education's National School Nutrition Programme.

isiZulu was included in the data sets that were collected to support IARPA/DARPA's[6] *Babel* and *LORELEI* (Low Resource Languages for Emergent Incidents) programs (Harper 2011, Strassel & Tracey 2016). These programs resulted in numerous investigations on the development of automatic speech recognition and spoken term detection capabilities in low-resource languages, many of which included isiZulu as an example language.

The National Centre for Human Language Technology (NCHLT) subsequently funded two projects to collect substantially larger speech and text data sets than those that were compiled during previous projects. The data collection efforts therefore went beyond the telephone-based, limited domain scope of the AST and Lwazi projects. This effort resulted in 11 speech corpora containing 50-60 hours of orthographically transcribed broadband speech per language and 11 text corpora of between 1.15 and 3.27 million words per language (Barnard et al. 2014, Eiselen & Puttkammer 2014). With the exception of a few domain specific data collection efforts (Davel et al. 2011, de Wet et al. 2011, de Wet et al. 2016), these remain the most extensive resources that are available for speech technology development in the country.

The majority of the projects mentioned in the previous paragraphs were associated with a specific institution in South Africa, whereas anybody from any language community (who adheres to Mozilla's code of conduct) can participate in the Common Voice project. Anybody who would like to can therefore contribute to the data collection effort and become involved in the local language technology community. The Common Voice project also has an international reach which means that the South African community stands to benefit from "lessons learnt" during localisation and data collection in other countries as well as the technical support provided on Common Voice community user groups. The rest of the paper describes how the Common Voice website was localised and presents the results of an initial attempt to harvest sentences from text data on the web.

3 Mozilla Common Voice platform

As was mentioned in Section 1, a language needs to meet two requirements to be launched on the Common Voice platform: 1) the website needs to be translated into the target language and 2) 5 000 sentences that are in the open domain need to be collected. Mozilla provides tools (via websites and community user groups) and guidelines to assist with both these processes. Their implementation in the current project is briefly described in the next two sections of the paper.

3.1 Translation

A number of service providers were requested to submit quotes for translating the Common Voice website from English into the 10 other official languages. A company with previous experience in localisation was identified as the best candidate to perform the translations. All translations were performed by the same company so that they could manage aspects like the standardisation of terminology between languages in the same manner for all languages.

Words and utterances were translated using Mozilla's translation tool, Pontoon[7], before being used to generate language specific web pages automatically. The resulting web pages were subsequently proofread by a second team of linguists. The feedback they provided ranged from remarks on lexical choice based on differences in intra-lingual geographical variation between the dialects of the translators and proofreaders (despite both parties being native speakers of the language) to the way in which Mozilla's technology utilized words to create automatic translations without taking the morphological makeup of a word into account. Amongst other things their comments indicated that Mozilla's tools do not make provision for the noun class agreement system in the Nguni languages, resulting in words translated in isolation not appearing correctly in sentences. In some cases the tools did not accommodate the length of words in more descriptive sentences where words are more morphologically complex.

Translators and proofreaders also found it difficult to perform the localisation because many of the languages do not currently have words for technological terminology such as *part-of-speech tagger* or *sentence builder* resulting in the choice between creating or establishing terminology which might alienate potential users of the website, or using existing English terminology which might create the perception that the entire website has not been localised.

3.2 Text collection

Although a number of curated text corpora collections already exist for all of the South African languages, there are several complications to using these corpora as example sentences for the Mozilla Common Voice project. Firstly, almost all of these corpora are distributed under CC-BY licenses, similar to those used by open source initiatives such as Wikipedia. This implies that only subsections of these data sets (typically less than 10% of the original article) qualify as CC0. Secondly, many of these data sets are either sourced from government documents, speeches, and websites, or from religious ma-

terial, such as the Bible. These texts represent very specific domains, subject matter, as well as writing style.

To mitigate these problems, we investigated the possibility of sourcing text data from other web sources that are commonly used in language technology development. For instance, Mozilla provides a set of text processing tools to harvest data from Wikipedia. The tools could not be used "as is" in this project, because the default English rule set only allows sentences with ASCII characters. However, most of the South African languages include diacritic markers encoded by UTF-8 characters. The rule set therefore had to be adapted to accept within-language UTF-8 characters but to reject irrelevant ones.

One of the Mozilla selection rules specifies that only three sentences may be copied from a Wikipedia page, but only if the article contains 10 or more sentences. Many articles in South African languages did not meet the 10-sentence limit and, as a result, no sentences could be harvested from them. Another limitation that became evident is the lack of lists of "disallowed words" in most of the languages. These lists are used to prevent possibly offensive words from appearing in the sentences. The text collected from Wikipedia was also verified using automatic Language Identification (LID) (Puttkammer et al. 2018, Hocking 2014)[8]. The verification revealed that many articles contain text in languages other than the target language. These sentences were discarded.

Although the Mozilla tools and Wikipedia could be used to obtain some data in a few languages, the current Wikipedia presence of the majority of the languages yielded less than a thousand sentences per language. Moreover, the text collection did not produce any isiNdebele sentences because, at the time of writing, the language did not have a presence on Wikipedia.

Other sources of web-based text were subsequently explored in an attempt to collect isiNdebele sentences as well as additional data for the other languages. These included the Leipzig Corpora Collection (LCC) (Goldhahn et al. 2012), OPUS (Tiedemann & Nygaard 2004) and the FLORES-200 (Goyal et al. 2022) data sets. The Mozilla tools were also used to collect sentences from these sources, but with the restriction that no more than 9.5% of any particular source was allowed to be harvested. The resulting selections were also verified using LID and the same rules for discarding unwanted characters were applied.

In addition to these pre-processing steps, the text was sentence separated and frequency lists were generated using CTexTools 2 (Puttkammer et al. 2018)[9]. Sentences or segments that did not include useful data (e.g. lines containing only telephone numbers or punctuation) as well as lines that did not constitute a well formed sentence (starting with optional punctuation or numbering, then a capital letter and ending with sentence ending punctuation) were removed from the sentence separated data. The sentences were also filtered to contain between three and fourteen words but with an absolute character limit of 99 [10]. Sentences including numerals were also removed according to the guidelines [11].

The frequency lists were then spell checked using commercially available spelling checkers [12] developed by the Centre for Text Technology at the North-West University in South Africa [13]. Using the spell checked lists, all remaining sentences were ranked according to the percentage correctly spelled words they contain and only sentences with more than 80% correctly spelled words were kept. To ensure better coverage of the languages, these sentences were then compared using the Levenshtein edit distance [14]. Only sentences with less than 70% overlap were included in the final set for each language.

After completing the above mentioned steps, only the Afrikaans and Setswana texts still contained more than 5 000 sentences. This was partially due to the fact that the initial corpora were relatively small. Another contributing factor is an overlap of up to 80% between the web-based corpora like LCC and OPUS. This observation seems to suggest that the

text collections were probably obtained form the same sources.

Searching for some of the terms in the Afrikaans list of disallowed words revealed that adding this type of filtering is essential to prevent offensive words and sentences from appearing in the sentences. A similar process also indicated that there is a strong presence of religious text in the harvested data, despite a concerted effort to avoid religious and government publications. Appropriate filters for these types of texts will therefore also have to be designed for each of the 10 languages under consideration before adding any text harvested from the web to the Common Voice platform.

4 Future work

Immediate next steps in the project will be to address the issues discussed in the previous section in order to reach the target of 5 000 CCo sentences per language. The South African governmental websites (*.gov.za) appear in all the official languages and the possibility to obtain additional text from this source will be investigated. Once the required number of sentences have been collected, the Common Voice websites will be ready for speech data collection to start. The project will be promoted as widely as possible with the aim to encourage language communities across the country to become involved. Hopefully these efforts will be successful to the extent that data collection can be followed by dedicated speech technology development workshops.

Notes

[1] https://www.undp.org/sustainable-development-goals, https://odi.org/en/publications/leave-no-one-behind-index-2019/

[2] https://commonvoice.mozilla.org/en

[3] https://creativecommons.org/share-your-work/public-domain/cc0/

[4] All resources that were generated with government grants are made freely available and are accessible via the South African Centre for Digital Language Resources' Resource Catalogue: https://repo.sadilar.org/.

[5] In the Nguni languages spoken in South Africa *lwazi* means knowledge or information.

[6] Intelligence Advanced Research Projects Activity/Defense Advanced Research Projects Agency

[7] https://pontoon.mozilla.org/projects/common-voice/

[8] https://hdl.handle.net/20.500.12185/350

[9] https://hdl.handle.net/20.500.12185/480

[10] https://discourse.mozilla.org/t/using-the-europarl-dataset-with-sentences-from-speeches-from-the-european-parliament/50184

[11] https://commonvoice.mozilla.org/sentence-collector/##/en/how-to

[12] https://spel.co.za/en/product/african_spelling_checkers/

[13] https://humanities.nwu.ac.za/ctext

[14] https://metacpan.org/pod/Text::LevenshteinXS

Acknowledgements

The localisation of the Mozilla Common Voice platform in South Africa is supported by the German Federal Ministry of Economic Cooperation and Development (BMZ), represented by the GIZ project FAIR Forward - Artificial Intelligence for All.

References

Adda, G., Choukri, K., Kasinskaite, I., Mariani, J., Mazo, H. & Sakriani, S., eds (2019), *Proceedings of the 1st International Conference on Language Technologies for All*, European Language Resources Association (ELRA), Paris, France.

Barnard, E., Davel, M. H., van Heerden, C., de Wet, F. & Badenhorst, J. (2014), The NCHLT Speech Corpus of the South African languages, *in* 'Proceedings of the 4th Workshop on Spoken Language Technologies for Under-resourced Languages', St Petersburg, Russia, pp. 194–200.

Barnard, E., Davel, M. H. & Van Huyssteen, G. B. (2010), Speech Technology for Information Access: A South African Case Study, *in* 'AAAI Spring Symposium Series', pp. 8–13.

Calteaux, K., de Wet, F., Moors, C., van Niekerk, D., McAlister, B., Sharma-Grover, A., Reid, T., Davel, M., Barnard, E. & van Heerden, C. (2013), Lwazi II Final Report: Increasing the impact of speech technologies in South Africa, Technical report, CSIR.

Davel, M. H., van Heerden, C., Kleynhans, N. & Barnard, E. (2011), Efficient harvesting of Internet audio for resource-scarce ASR, *in* 'Proceedings of Interspeech', International Speech Communication Association (ISCA), Florence, Italy, pp. 3153–3156.

de Wet, F., Badenhorst, J. & Modipa, T. (2016), 'Developing Speech Resources from Parliamentary Data for South African English', *Procedia Computer Science* **81**, 45–52.

de Wet, F., de Waal, A. & van Huyssteen, G. B. (2011), Developing a broadband automatic speech recognition system for Afrikaans, *in* 'Proceedings of Interspeech', International Speech Communication Association (ISCA), Florence, Italy, pp. 3185–3188.

Eiselen, R. & Puttkammer, M. J. (2014), Developing Text Resources for Ten South African Languages, *in* 'Proceedings of Language Resource and Evaluation (LREC)', Reykjavik, Iceland, pp. 3698–3703.

Goldhahn, D., Eckart, T. & Quasthoff, U. (2012), Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages, *in* 'Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC)'.

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F. & Fan, A. (2022), 'The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation', *Transactions of the Association for Computational Linguistics* **10**, 522–538.

Harper, M. P. (2011), 'Data Resources to Support the Babel Program Intelligence Advanced Research Projects Activity (IARPA)'.

Hocking, J. (2014), Language identification for South African languages., *in* 'Proceedings of the Annual Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)', PRASA.

Joshi, P., Santy, S., Budhiraja, A., Bali, K. & Choudhury, M. (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', Association for Computational Linguistics, pp. 6282–6293.

Kuun, C. (2012), Development of a telephone-based speech-driven information service for the

South African Government, Technical report, CSIR.

Puttkammer, M., Eiselen, R., Hocking, J. & Koen, F. (2018), NLP Web Services for Resource-Scarce Languages, *in* 'Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)', Association for Computational Linguistics, pp. 43–49.

Roux, J. C., Louw, P. H. & Niesler, T. (2004), The African Speech Technology Project: An Assessment, *in* 'Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)', European Language Resources Association (ELRA), Lisbon, Portugal, pp. 93–96.

Strassel, S. & Tracey, J. (2016), LORELEI Language Packs: Data, Tools, and Resources for Technology Development in Low Resource Languages, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', European Language Resources Association (ELRA), Portorož, Slovenia, pp. 3273–3280.

Tiedemann, J. & Nygaard, L. (2004), The OPUS corpus - parallel and free: `http://logos.uio.no/opus`, *in* 'Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)', European Language Resources Association (ELRA), Lisbon, Portugal.

Titmus, N., Schlünz, G. I., Louw, J. A., Moodley, A., Reid, T. & Calteaux, K. (2016), Lwazi III Project Final Report: Operational Deployment of Indigenous Text-to-Speech Systems, Technical report, CSIR.

A Lexical database of Malagasy adjectives

Joro Ranaivoarison
Department of Malagasy, University of Antananarivo
jororanaivo@llm-u-ank.mg

Abstract

This paper deals with an electronic resource under construction. The objective is to construct a lexical database of Malagasy adjectives. Malagasy is an agglutinative language of Austronesian origin, spoken in the African island of Madagascar. The method used to construct the resource is adapted from the approach of Gross (1989) to electronic dictionaries. The content of the resource is based on linguistic analysis and encoded so as to be used by language-processing software. However, care has been taken that all linguistic information is easily readable and updatable. The resource allows for morphological analysis and generation of adjectives, removing obstacles to the construction of computer applications to process the Malagasy language. The originality of this paper also comes from our proposal of a distinction between adjectives in the usual sense and adjectival forms of other parts of speech.

Keywords: lexical database, Malagasy, adjective, inflection, derivation

1 Introduction

This research takes place in the context of the construction of a general resource for Malagasy language, and focuses on adjectives. This resource is under construction and must be revised and enriched.

The database is designed to be used by language-processing programs. This goal requires a high level of precision and formalization. We borrowed methodological principles from similar successful projects and we used Unitex/GramLab (Paumier, 2003), an open-source, freely available platform of language processing that also offers functionality for constructing language resources.

As compared to traditional grammar and previous work by linguists on adjectives, we distinguish adjectives in the usual sense from adjectival forms of other parts of speech (in this case, nouns).

The next section provides general information about adjectives in Malagasy. Section 3 sums up the approach of Gross (1989) to electronic dictionaries. In Section 4, we describe how (morpheme-internal) allomorphy is represented in the database. Section 5 is about the dictionary of adjectival lemmas in the usual sense, and Section 6 is about the resources that account for adjectival forms of nouns. We report the experiments performed to test the resources in Section 7. The last section contains concluding remarks.

2 Adjectives in Malagasy

Language-independent definitions of adjectives, e.g. "a term used in grammatical classification of words to refer to the main set of items which specify the attributes of nouns" (Crystal, 2008: 11), are suitable for Malagasy, but not precise enough, for example, to distinguish adjectives from verbs with the *m-* prefix. For that matter, we follow Rajaona (1972: 404) and Ralalaoherivony (1995) who draw the line by using the following criterion: verbs have forms in the circumstantial voice, while adjectives do not. Refer to Rajaona (1972: 521), Keenan & Polinski (1998: 609), Dalrymple et al. (2005) about circumstantial voice. So, *mandèha* "walk", *mandrày* "take" are verbs because the circumstantial forms *andehánana, andráisana* are in use; and elements as *makòtroka, màtotra* are adjectives because the circumstantial forms **akotróhana, *atórana* are not attested. (Graphical accents in Malagasy are optional and indicate stress.)

This section reports on the inflection of adjectives and on the distinction between derived and inflected adjectives that stem from nouns.

2.1 Inflection of adjectives

Like verbs, some adjectives receive inflectional markers of grammatical tense in Malagasy (cf. Rajaonarimanana, 1995: 64).

Most adjectives combine with *ho* to express future tense, hence *màro* "is/was numerous" /*ho màro* "will be numerous", *mànta* "is/was raw"/*ho mànta* "will be raw", etc. This marker of the future tense is not an inflectional morpheme because it can be separated from the adjective by other words.

However, other adjectives show morphological alternation associated with tense, e.g. *m-a-fàna* (PRS-ADJZ-heat) "is hot", *n-a-fàna* (PAST-ADJZ-heat) "was hot" and *h-a-fàna* (FUT-ADJZ-heat) "will be hot". Among the about 2,115 adjectives studied for this paper, about 365 take the *m:n:h* alternation.

Traditional grammarians as Malzac (1950), Rajemisa (1969), and linguists like Rajaona (1972), Rajaonarimanana (1995) consider that *m-, ma-, mi-, man-, -ina, -ana* are adjective-forming affixes. Among these, we choose to segment *ma-, mi-, man-* further into the tense markers *m-, n-, h-* and the adjective-forming affixes *a-, i-, an-*, whenever the morphological alternation *mi-, ma-, man-/ni-, na-, nan-/hi-, ha-, han-* correlates with the difference of tense.

Some adjectives in Malagasy take imperative suffixes (cf. Rajaona, 1972; Catz & Catz, 2017) such as *-a* in *m-a-heréz-a* (PRS-ADJZ-power-IMP) "be powerful" vs. *m-à-hery* (PRS-ADJZ-power) "powerful", *m-a-zoto-a* (PRS-ADJZ-diligence-IMP) "be industrious" vs. *m-a-zòto* (PRS-ADJZ-diligence) "industrious". For now, the imperative mood is not taken into account in the resource.

2.2 Derived or inflected adjectives

Many Malagasy adjectives are derived from nouns. As is usual in derivational morphology, the meaning or the syntax of the derived adjective is not entirely predictable. In Table 1, for example , the base nouns *kòtroka*, *hìdy*, *jèmby* have the following derivations: *m-a-kòtroka, m-a-hìdy, jembé-na*.

Table 1: Derived forms with unpredictable meaning or syntax

Base noun	Derived adjective	English gloss
kòtroka "thunder"	*m-a-kòtroka*	warm
hìdy "lock"	*m-a-hìdy*	selfish
jèmby "confusion"	*jembé-na*	very dark

Other adjective-forming affixes have been described in Malzac (1950), Rajemisa (1969) and Rajaona (1972)[cf. Subsection 5.5 below].

In Malagasy, however, the grammatical relation between nouns and adjectives can also fall under inflectional morphology. For about 5% of the adjectives we studied in this research, the meaning and syntax of the denominal adjective are entirely predictable based on the noun stem and the morphological process involved, as in Table 2.

Table 2: Inflected forms with predictable meaning and syntax

Base noun	Inflected adjective	English gloss
hànitra "fragrance"	*m-ànitra*	fragrant, aromatic
tànjaka "strength"	*m-a-tànjaka*	strong
fàika "dregs"	*faiká-na*	dreggy

Such pairs are so regular that the process can be regarded as inflectional, even if the base form and the derived form belong to distinct parts of speech. In other words, these adjectives can be considered as inflected adjectival forms of the corresponding nouns, just like participles are inflected adjectival forms of verbs in English: on the one hand, they may behave as adjectives (cf. the term 'participial adjective' used for example by Kennedy & McNally (1999)), but on the other hand they belong to verb conjugation.

The resource includes derived adjectives such as *m-a-kòtroka* (PRS-ADJZ-thunder) "warm", adjectival inflected forms of nouns like *m-ànitra* (ADJZ-fragrance) "fragrant", and base adjectives like *àvo* (ADJ) "high".

The distinction between derived and inflected adjectives is an inescapable reality of Malagasy, but the main affixes serve both as derivational and inflectional:

- *i-*, in *m-i-kodiadìa* "big and fat (of a child)", from *kodìa* "wheel", is derivational, but in *m-i-kitoantòana* "rough, uneven, craggy", from *kitoantòana* "uneven ground, rough place", *i-* is inflectional;
- as for *a-* in *m-a-tètika* "frequent", from *tètika* "ornamental scarification; cutting up small pieces", it is derivational, but in *m-a-fàna* "hot", from *fàna* "heat", *a-* is inflectional;
- as for *m-* in *m-èndrika* "fit, proper, worthy", from *èndrika* "face, likeness, image", *m-* is derivational, but in *m-ànitra* "fragrant", from *hànitra* "fragrance", *m-* is inflectional.

Consequently, during the construction of the resource, we face difficult decisions in classifying denominal adjectives as derived or inflected, especially when the prefix forming the adjective is *m-*. In such cases, we analyse the adjective as an inflected form of a noun only when we identify a pair of syntactic constructions such as *Misy hànitra* N_0 "N_0 has fragrance" = *Mànitra* N_0 "N_0 is fragrant", where *mìsy* "there is" or *mànana* "have" is a support verb (Ranaivoson, 1996; Lakoarisoa et al., 2011; Jaozandry, 2014; Hamitramalala, 2017), and N_0 is an accepted subject noun. In the case of *m-ày* "on fire", *m-àfy* "hard", respectively from *hày* "burning", *hàfy* "hardship", there are no such pairs of syntactic constructions, since the nominal construction is not in use:

Mày N_0 "N_0 is on fire"
**Mìsy/Mànana hay* N_0

Màfy N_0 "N_0 is hard"
**Mìsy/Mànana hàfy* N_0

Thus, we encoded them as derived adjectives in the resource. Such lexicological decisions border on the arbitrary, but formal criteria are the best way we know to make them reproducible.

Verbs with a resultative prefix, e.g. *mahatalànjona* "amazing", *mahavàriana* "stunning" are considered as verbal forms of *talanjona* "amazed", *varìana* "stunned" (cf. Rajaona, 2004:58) and encoded in our database in the framework of conjugation (Ranaivoraison et al., 2013).

3 Electronic dictionaries

The method used to construct the resource for adjectives of Malagasy is adapted from the approach of Gross (1989) to electronic dictionaries. This approach recommends several methodological safeguards.

First, for lexical databases to be usable by programs, all data must be explicit. This situation contrasts with that of dictionaries for human readers, where some information may remain implicit, since readers rely on their linguistic proficiency to infer it.

Next, lexicological and lexicographical decisions are based on the observation of a sufficiently large number of lexical entries. Entries are systematically inventoried and decisions are based on this inventory, not on sporadic observations on a limited sample of entries, a practice that would be more likely to necessitate revisions of these decisions.

The resources must be readable, so that they can be updated.

Finally, modes of inflection are defined explicitly, so that the inflected forms of a stem can be generated automatically. A mode of inflection is the set of morphological changes underwent by a stem to generate its inflected forms. Several lexical entries can share the same inflectional mode, as *ring* and *sing* in English. The method requires that inflectional modes are defined independently of one another, in order to avoid constructing a hierarchy of general rules and exceptional rules, since such hierarchies are usually complex to maintain later, as the database undergoes updates. As each inflectional mode is independent, updating one does not require updating others. In consequence, we assign an identifier to each inflectional mode and we mark

in each entry of the dictionary the identifier of the type of inflection applicable. Thus, knowing if a rule is applicable to a lexical entry does not require applying it. This contrasts with Two-level morphology (Koskenniemi, 1983), where rule scope is encoded in rules, so that knowing if a lexical entry is affected by a change in a rule requires applying both versions of the rule. In addition, two-level rules are ordered, so that knowing if a lexical entry is affected by a change in rule order requires applying both versions of all rules. These features are practical obstacles to updates, corrections and extensions of two-level databases.

Thus, our approach encodes the generation of inflected forms from their stems, providing a formal link between them, e.g. between *sing* and *singing* in English. However, the approach does not do the same between derivatives and their bases, as *speak* and *speech*, since semantic and syntactic irregularities reduce the potential applications of such generation. Thus, derived words such as *mèndrika* "fit, proper, worthy", *m-i-kodiadìa* "big and fat (of a child)" are encoded as stems.

Gross' approach, devised for inflectional languages, has been extended to agglutinative languages (Berlocher et al., 2006; Ranaivoarison et al., 2013) by distinguishing two levels of morphological changes:

- morpheme-internal allomorphy, e.g., the noun *sómotra* "beard" becomes *somór* immediately before some suffixes,

- affixations, e.g., *-ina*, a morpheme of formation of adjectival forms, can be suffixed to *somór*, giving *somórina* "bearded".

Through the two steps corresponding to these two levels, the inflected form *somòrina* can be generated from the lemma *sómotra*, or conversely, the lemma can be recognised from the inflected form.

The Unitex open-source platform of language processing (Paumier, 2003) is compatible with resources devised according to this approach and encoded in the DELA format (Gross 1989: 8). Unitex performs top-quality inflection, compression and lookup (Neme, Paumier, 2019) of lexical databases encoded in this format. In contrast, the Text encoding initiative's dictionary encoding formats are mainly designed for human-oriented dictionaries (TEI Consortium, 2022) and do not address the processing of lexical databases.

In the electronic-dictionary approach, lexical databases mainly cover parts-of-speech and inflection. This contrasts with the lexicon-grammar approach, also proposed by Gross (Elia, 1978), which also investigates the applicability of syntactic operations. Due to this difference, sense distinctions are more fine-grained in the latter approach (Laporte, 1991).

4 Allomorphy

This section provides information about morpheme-internal allomorphy and how it is encoded so that lexical variants of stems can be produced automatically. We list the phenomena that affect stems and we describe graphs that encode allomorphy.

4.1 Phenomena of allomorphy

Immediately before or after affixes, some stems do not vary, as in *èrika* "drizzling rain"/ *m-èrika* "drizzly, misty"*, dìo* "cleanliness, purity" /*m-a-dìo* "clean, clear, pure", but most do. They can undergo 5 types of variation:

- prosodic alternations as in *sòmotra*/*somór* (the accent shifts from the first *ò* to the second *ó*)
- insertion of a letter as in *safòfoka*/*tsafòfoka* (insertion of *t* before the stem in the inflected form *mantsafòfoka*)
- substitution of a letter *hàtsiaka*/*gàtsiaka* (substitution of *g* in the inflected form *mangàtsiaka* for *h* in the stem)
- extension as in *nòfo*/*nofós* (insertion of *s* in the inflected form *nofósana*)
- deletion of a letter as in *vorètra*/*orètra* (deletion of the letter *v* of the stem in the inflected form *mamorètra*) or in *sòmotra* /

somór (deletion of the *t* and *a* of *sòmotra* in the inflected form *somórina*)

All allomorphic stems of adjectives undergo one or several of the phenomena above in their lexical variants.

For a deeper study of the phenomena affecting the stems in Malagasy, refer to Rajaona (2004).

4.2 Encoding allomorphy with graphs

To encode the lexical variants of stems, we attach to each entry a code that identifies accurately the applicable variations. These variations are encoded in the form of a graph like that of Fig. 1, which produces automatically the lexical variants of stems like *pànda* "freckles", *kilèma* "deformity", *kìbo* "belly".

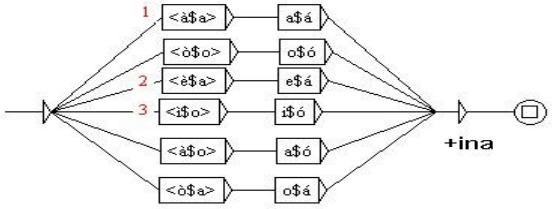

*Figure 1: Graph of allomorphy '**0v**'*

Path 1 encodes *pànda* > *pandá* in *pandáina* "freckly, sunburnt"; path 2, *kilèma* > *kilemá* in *kileméina* "maimed"; and path 3, *kìbo* > *kibó* in *kibóina* "big-bellied". This technique of lexical marking facilitates updates. The '**0v**' identifier is the name of the graph and is attached to those words to encode these prosodic alternations. So, several lexical entries such as *tràtra* "breast", *tsikòko* "scabs" share the same inflectional mode '**0v**' (identifier) to produce automatically lexical variants as *tratrá, tsikokó*.

Other lemmas have *-ka, -tra, -na* endings, and their final *-a* is missing in most of their allomorphs: the corresponding graphs specify the deletion of this vowel.

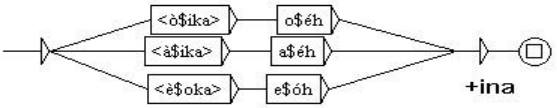

*Figure 2: Graph of allomorphy '**1v**' for lemmas in* -ka.

For example, the '**1v**' graph in Fig. 2 deletes the ending *-ka* and inserts a letter *h*, in addition to the prosodic and/or vocalic alternations, encoding variations such as *dònika* > *donéh* in *donéhina* "suffering from mumps", *fàsika* > *faséh* in *faséhina* "sandy", *tsèroka* > *tseróh* in *tseróhina* "dirty, scurfy".

If the lemma contains inflectional prefixes, the graphs also strip them off.

Thus, graphs of allomorphy for stems in Malagasy are divided in 4 types: for those with *ka, tra,* or *na* endings, and for those with none of these endings. In addition, they specify some of the five phenomena identified in Subsection 4.1. Thirty graphs of allomorphy are used for about 2,115 adjectives.

We will now describe the other resources for lexical entries of adjectives, and then those for adjectival forms of nouns.

5 Encoding lexical entries of adjectives

Base adjectives such as *àntitra* "old", *àvo* "high", and derived adjectives such as *makìkitra* "determined", *makòtroka* "warm" are all encoded in the lexical database as adjectival entries without distinction.

We report on the different components of this resource: lexical entries and graphs of affixation, and we mention the main inflectional and derivational morphemes of adjectival entries.

5.1 Lexical entries

Our resource contains about 2,000 adjectival lemmas (Fig. 3). Entries and morphological codes are separated with a comma. The letter "A" is the morphological code of adjectives. The code in parentheses indicates the graph of allomorphy applicable to the entry, and the name "ad1"

indicates the graph of affixation for the words that take the *m-:n-:h-* morphological alternation, as in *malàdy* "is quick to hear": *nalády* "was quick to hear": *halády* "will be quick to hear".

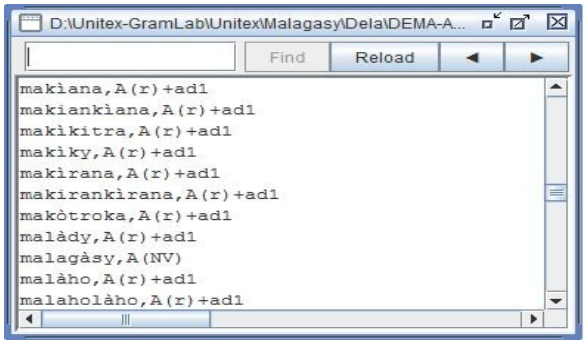

Figure 3: Resource of adjectival entries.

5.2 Graphs of affixation

Graphs of affixation specify which combinations of grammatical and lexical morphemes make up inflected forms of adjectives. In the present state of the lexical database, the graphs of affixation take into account the tense prefixes *m-:n-:h-*, but not yet affixes for imperative.

Take for example the lemmas *marènina* "deaf", *makàka* "spacious". The *m-* in the beginning is the prefix of present tense, so the stems are *arènina, akàka,* which are generated by the graph of allomorphy identified in their entries. The "ad1" graph of affixation (Fig. 4) specifies that this stem takes the three tense prefixes.

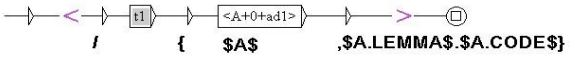

Figure 4: Graph of affixation "ad1".

In this graph, the '**t1**' box is a call to a subgraph that represents the *m-:n-:h-* morphological alternation. The **<A+0+ad1>** box stands for the relevant variant of the stem, i.e. *arènina, akàka,* among others.

Among the about 2,000 adjectival entries, about 505 begin with *m*, about 300 of which take the *m:n:h* alternation. Adjectives without the *m-:n-:h-* alternation, such as *mòana* "speechless" (**nòana* and **hòana* are not in use), form a syntactic future with *ho*, as in *ho mòana* "will be speechless".

5.3 Inflectional and grammatical codes

The codes of parts of speech used in this part of the database are:

- A for adjective
- T for tense marker

Inflectional codes are:

- r, p, f respectively for present, past, future tenses
- n for indicative mood.

5.4 Main inflectional morphemes of adjectival entries

Table 3 contains the main inflectional morphemes used in this part of the database.

Table 3: Main inflectional morphemes of adjectives

	Affix	Example	Gloss
Prefixes of tense (T)	*m-*	*madio*	clean (present)
	n-	*nadio*	clean (past)
	h-	*hadio*	clean (future)

5.5 Derivational affixes of adjectives

Table 4 shows different affixes of adjectives generally considered as derivational. The words containing those affixes are not numerous. In the resource, we analyse them as derivatives and encode them as adjectival lemmas. Since our database does not attempt to link them to their bases (cf. Section 3), these affixes have no formal existence as such in the database: the affix/base segmentation is not encoded.

In some denominal adjectival lemmas in *m-*, the initial *m-* is not present in the base noun, as in *màmy* "sweet" from *hàmy* "sweetness", *màty* "dead" from *fàty* "death", but it does not take the *m:n:h* tense alternation. In these entries, we analyse *m-* as a derivational adjectivizing prefix, just like those of Table 4, and we do not encode the affix/base segmentation.

Table 4: Derivational affixes of adjectives

Affix	Example	Gloss

ba-	*bakaka*	badly plaited, as mats
do-	*dorehitra*	very red, scarlet
fo-	*forehitra*	consumed
faha-	*faharoa*	second
fa-	*farofy*	sickly
ka-	*kaozatra*	very lean
ki-	*kiboribory*	round
ki-…-ina	*kibotaina*	plump
ko-	*kosesy*	frequent
sa-	*saresaka*	talkative
so-	*somatroka*	drab
ta-	*takariva*	about dusk
tan-…-ana	*tandemena*	faint
to-	*tolantsika*	arqued
tsa-	*tsatselika*	agile
tsi-	*tsilotidotika*	dirty
va-	*varozaka*	weak, exhausted
-om-	*somariaka*	glad
-il-	*kilitika*	extremely small
-ir-	*kiritika*	extremely small

6 Adjectival forms of nouns

In our lexical database, adjectival inflected forms of nouns as *vintánina* "who has a destiny", from *vìntana* "destiny", or *mazòto* "diligent, industrious", from *zòto* "diligence", are encoded in a dictionary of nouns, in the form of resources that allow for segmenting and generating these adjectival forms. For example, the noun *lòto* "filth, dirtiness" receives the inflectional prefix forming adjectives *a-*, and the tense prefixes *m-*, *n-* or *h-*, hence *malòto* "is filthy, dirty", *nalòto* "was filthy, dirty", *halòto* "will be filthy, dirty".

We report in this section on the resources for these forms: lexical entries and graphs of affixation, and we list the relevant inflectional morphemes.

6.1 Lexical entries

The resource contains about 400 entries of nouns (Fig. 5), 115 of which have adjectival inflected forms. The code "N" is for 'noun' and the codes in parentheses identify the graphs of allomorphy. The graphs of allomorphy are those described in Subsection 4.2. The codes "A6", "A2" or "A3" are for the graphs of affixation. For example, nouns with the code A3, such as *lòto,* receive the *m-*: *n-*: *h-* morphological alternation of tense and the prefix of formation of adjectives *a-*, hence *malòto*, which is thus segmented as *m-a-lòto*, where the lemma is represented by the nominal stem *lòto*. The 'A3' graph encodes this segmentation in three morphemes and formally represents *a-* as an inflectional prefix.

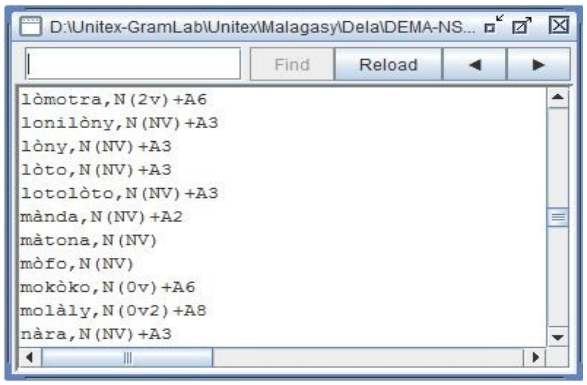

Figure 5: Resource for adjectival inflected forms of nouns.

This contrasts with *marènina* "deaf", analysed as *m-arènina* in Subsection 5.2: as *marènina* is an adjectival lemma, it does not contain any inflectional prefix of formation of adjectives, and the first *a* is part of the stem. Thus, graphs of affixation for adjectival forms of nouns are different from those for adjectival entries.

6.2 Graphs of affixation

We inventory thirteen graphs for adjectival forms of nouns.

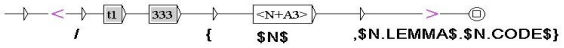

Figure 6: Graph of affixation 'A3'.

The 'A3' graph (Fig. 6) describes the combinations of affixes that make up adjectival forms such as *m-a-dìty* "gummy, resinous", *m-a-sìra* "salty" and their own inflected forms. The '**t1**' box represents the *m-:n-:h-* morphological alternation, the '**333**' box contains the prefix *a-* forming adjectives (PAdj) and the **<N+A3>** box

indicates to which nouns the graph is applicable, here *dìty* "gum, resin", *sìra* "salt", among others.

6.3 Grammatical codes

The additional codes of parts of speech used in this part of the database are:

- N for 'noun'
- PAdj for 'prefix forming adjectives'
- SAdj for 'suffix forming adjectives'.

6.4 Inflectional morphemes of adjectives

About 65 of the nominal entries that we studied produce adjectival forms that take the *m-:n-:h-* alternation. Table 5 contains the additional inflectional morphemes used in this part of the database.

Table 5: Inflectional morphemes of adjectives

	Affix	Example	Gloss
Adjectivizing prefixes (PAdj)	*m-*	*màizina*	dark
	i-	*mimànda*	having defences
	a-	*matànjaka*	strong
	an-	*mangàtsiaka*	cold
	am-	*mamirifiry*	cold
Adjectivizing suffixes (SAdj)	*-ina*	*fasebina*	sandy
	-ana	*nofosana*	fleshy
	-na	*faikána*	dreggy

The inflectional adjectivizing prefix *m-* occurs in words as *màizina* "dark" from *àizina* "darkness", *mànitra* "fragrant" from *hànitra* "fragrance": in these forms, it does not take the role of tense prefix. They form a syntactic future with *ho*.

7 Tests

In a novel by Clarisse Ratsifandrihamanana, a well-known Malagasy writer, which has 2,400 sentences and about 55,600 tokens, we recognize with our database about 380 unique forms of adjectives. The representation of the result with Unitex is different for base or derived adjectives and for adjectival forms of nouns.

Two experiments are presented in this section: segmentation and generation.

7.1 Segmentation

In this experiment, the Unitex platform used the lexical database and the graphs to segment text. The segmentation of *mareforèfo* "a bit fragile", a derived adjective that accepts the *m-:n-:h-* morphological alternation, is presented in Fig. 7.

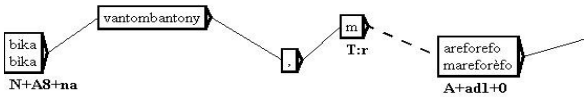

Figure 7: Representation of a derived adjective with a tense marker.

The two boxes connected by a broken line represent 1) *m-*, recognized as a tense marker (T) of present tense (r), and 2) *areforefo*, labeled **A+ad1+0**, which means that it is an adjective and lists two features; this form is attached to the lemma *mareforèfo*.

An adjectival inflected form of a noun is segmented as shown in Fig. 8.

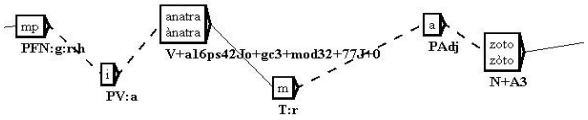

Figure 8: Representation of an adjectival form of a noun.

In *mpianatra mazòto* "industrious student", *mazòto* "diligent, industrious" is an adjectival inflected form of *zòto* "diligence". Unitex shows the morpheme of tense *m-* (present), the morpheme forming adjectives *a-* (Padj) and the nominal stem (N+A3).

7.2 Generation of adjectives

Unitex has a generator of words that lists the adjectival inflected forms specified by the database (Fig. 9). Each semicolon begins a new bundle of inflectional features. The tense codes **:r**, **:p** and **:f** are mutually exclusive because a word cannot be in the same interpretation at two different grammatical tenses, but the other code **n** qualifies the same interpretation as the tense preceding them. Thus, **A+A8+na:rn:pn** means that *bikána* "well

formed" can be in the present or in the past tense, but in both cases it is in the indicative.

```
bikána,bìka.A+A8+na:rn:pn
molaléna,molàly.A+A8+na:rn:pn
seréna,sèry.A+A8+na:rn:pn
sorisoréna,sorisòry.A+A8+na:rn:pn
teténa,tèty.A+A8+na:rn:pn
nofósana,nòfo.A+A7+ana:rn:pn
ranjóana,ranjo.A+A7+ana:rn:pn
rohánana,ròhana.A+A7+ana:rn:pn
sandríana,sàndry.A+A7+ana:rn:pn
tambavíana,tambàvy.A+A7+ana:rn:pn
vatovatóana,vatovàto.A+A7+ana:rn:pn
vatóana,vàto.A+A7+ana:rn:pn
vodíana,vòdy.A+A7+ana:rn:pn
váinana,vày.A+A7+ana:rn:pn
donéhina,dònika.A+A6+ina:rn:pn
fasipaséhina,fasipàsika.A+A6+ina:rn:pn
faséhina,fàsika.A+A6+ina:rn:pn
```

Figure 9: Generation of adjectives

8 Conclusion

We described a lexical database of Malagasy adjectives under construction. The content of the resource is based on linguistic analysis, bearing in mind relevant methodological safeguards. We make a distinction between adjectives in the usual sense and adjectival forms of nouns. The database can be used by language-processing software. The resource allows for morphological analysis and generation of adjectives. It is not limited to the adjectival forms occurring in a corpus of texts, but takes into account our competence as a native speaker and grammatical tradition.

This research has several potential applications. Its results can be used in grammar textbooks to describe the inflection of Malagasy adjectives. Linguists can use them to launch queries for grammatical configurations containing adjectives, e.g. noun phrases. Finally, lexical databases remove obstacles to the construction of computer applications to process languages.

Acknowledgements

We address our special thanks to the reviewers for their relevant comments and suggestions for the clarity of our paper.

References

Andriamise, L, Ranaivoson, JF, Rakotoalison, SF 2011, "Les locutions support en malgache. Le cas de *misy azy*," *Lexis and Grammar*, University of Cyprus, pp. 21-28.

Berlocher, I, Huh, HG, Laporte, E, Nam, JS 2006, "Morphological annotation of Korean with directly maintainable resources", *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa (Italy).

Catz, I & Catz, S 2017, *Standard Malagasy Grammar & List of 410+ Common Verb Conjugation,* edited by Kimmerling Razafindrina, Fetra Marc Humbert Rahajason and Vololona Fenohaja.

Crystal, D 2008, *A dictionary of linguistics and phonetics,* UK, Blackwell.

Dalrymple, M, Liakata, M & Mackie, L 2005, "A Two-level Morphology of Malagasy", *Proceedings of the 19th Pacific Asia Conference on Language, Information and Computation*, pages 83–94, Taipei, Taiwan, R.O.C, Institute of Linguistics, Academia Sinica.

Elia, A 1978 "Pour un lexique-grammaire de la langue italienne : les complétives objet", *Lingvisticae Investigationes* 2 (2), pp. 233-276.

Gross, M 1989, "La construction de dictionnaires électroniques", *Annales des Télécommunications*, vol. 44 (1-2), pp. 4-19.

Hamitramalala, R 2017, *Vers une typologie des collocations à verbe support en malgache*, PhD, Université de Montréal.

Jaozandry, M 2014, *Les prédicats nominaux du Malgache. Étude comparative avec le français*, PhD, Université Paris-Nord.

Keenan, E & Polinski, M 1998, "Malagasy (Austronesian)", *The handbook of Morphology,* ed.

Andrew Spencer and Arnold M. Zwicky, New Jersey, John Wiley & Sons.

Kennedy, Ch & McNally, L 1999, "From Event Structure to Scale Structure : Degree Modification in Deverbal Adjectives", Tanya Matthews and Devon Strolovitch (eds), *9th Semantics and Linguistic Theory Conference*, pp. 163-180.

Koskenniemi, K 1983, *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. PhD, University of Helsinki.

Laporte, E 1991, "Separating Entries in Electronic Dictionaries of French", *Sprache - Kommunikation - Informatik. Akten des 26. Linguistischen Kolloquiums, Poznan 1991*, J. Darski and Z. Vetulani eds., Tübingen: Max Niemeyer, pp.173-179.

Malzac, RP 1950, *Grammaire malgache,* Paris, Société d'éditions géographiques, maritimes et coloniales.

Neme, A & Paumier, S 2019, "Restoring Arabic vowels through omission-tolerant dictionary lookup", *Language Resources and Evaluation* 54, pp. 487-551.

Paumier, S 2003, *Unitex manual*, University of Marne-la-Vallée, Paris.

Rajaona, S 1972, *Structure du malgache*, Fianarantsoa, Ambozontany.

Rajaonarimanana, N 1995, *Grammaire moderne de la langue malgache,* Paris, L'Asiathèque.

Rajemisa-Raolison, R 1969, *Grammaire malgache,* Fianarantsoa, Ambozontany.

Ralalaoherivony, BS 1995, *Lexique-grammaire du malgache. Constructions adjectivales,* Thèse de doctorat, Université Paris 7.

Ranaivoarison, J, Laporte, E & Ralalaoherivony, BS 2013, "Formalization of Malagasy conjugation", *Language and Technology Conference*, Poznań (Poland).

Ranaivoson, JF 1996, *La nominalisation en malgache. Étude des formes* manao N, PhD, Université d'Antananarivo.

Richardson, J 1885, *A new Malagasy-English Dictionary,* Antananarivo, The London Missionary Society.

TEI Consortium, eds. (2022) *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* 4.4.

An overview of Sesotho BLARK content

Sibeko, Johannes
Nelson Mandela University
johannes.sibeko@mandela.ac.za

Setaka, Mmasibidi
South African Centre for Digital Language Resources
mmasibidi.setaka@nwu.ac.za

Abstract

This article overviews digital language resources available for Sesotho, an official language of South Africa. The South African Center for Digital Language Resources (SADiLaR) repository is used as a reference as it is the official host of various language resources for South African languages. A total of 18 written resources are identified from the repository, and a further 16 spoken resources are identified. Finally, a total of 45 applications and modules were identified. Findings indicate that the majority of applications and modules available for Sesotho are in fact general resources aimed at all eleven official South African languages. Furthermore, the available resources indicate an inclination to the development of entry level, basic language resources and an absence of middle and higher resources with functionalities such as semantic analyses for written resources and prosody prediction for spoken resources. The study is hindered by the dearth of resource specific evaluations and related research and exacerbated by the absence of some of the resources on the repository.

Keywords: Sesotho, BLARKs, Written resources, Spoken Resources, Digital language resources

1 Introduction

There is a growing interest in Human Language Technologies (HLTs) for low-resourced languages (LRLs) (Strassel & Tracey 2016). Accordingly, a number of HLT audits have been conducted on South African official languages (Grover et al. 2010, 2011, Moors, Wilken, Calteaux & Gumede 2018, Moors, Wilken, Gumede & Calteaux 2018). The language audits are aimed at two objectives that are (i) determining resources that need to be developed, and (ii) opportunities for multidisciplinary research. South Africa currently recognizes eleven official languages, namely, Afrikaans, English, isiZulu, isiXhosa, Siswati, Xitsonga, Tshivenda, isiNdebele, Setswana, Sepedi and Sesotho. In this article, we pay special attention to digital language resources available for Sesotho, a Bantu language that forms part of the bigger Sotho-Tswana group with Sepedi and Setswana (Riep 2013, Van Heerden et al. 2010, Nkolola-Wakumelol et al. 2012, Mojela 2016). Additionally, Sesotho is one of the official languages in Lesotho, and an officially recognised language in Zimbabwe (Ndlovu 2011, 2013, Kadenge & Mugari 2015, Wissing & Roux 2017). Sesotho has developed as both a spoken and written language (Moeketsi 2014, Koai & Fredericks 2019), and is used in a variety of domains.

Of the eleven official languages of South Africa, two languages, namely: Afrikaans and South African English, have the most digital language resources followed by Setswana, Sepedi, isiZulu, isiXhosa and Sesotho in no particular order (Moors, Wilken, Calteaux & Gumede 2018, Moors, Wilken, Gumede & Calteaux 2018). This reality is propelled by the lack of investment into the remaining nine indigenous languages. This image was illustrated at the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (Pretorius et al. 2014). Nonetheless, although there are digital language resources in Sesotho, the language remains marginalised and under-resourced as it is yet to have enough high quality supervised data (Koai & Fredericks 2019, Magueresse et al. 2020). As a result, multiple studies identify Sesotho as a low-resourced language (Mahloane & Trausan-Matu 2015, Chiguvare & Cleghorn 2021, Hanslo 2021, Sibeko & Van Zaanen 2022*a*). Moors, Wilken, Gumede & Calteaux (2018) go as far as identifying it as a severely under-resourced language. That is, al-

though it is taught as a home language and an additional language in both South Africa and Lesotho, it persists to be less studied for computerization, and thus remains less privileged and of low density (Cieri et al. 2016, Magueresse et al. 2020).

Although there have been language audits of all eleven official languages of South Africa, as far as we could ascertain for this article, no study has paid in depth attention to Sesotho digital language resources. Ensuingly, this article seeks to describe basic digital language resources, modules, and applications used in written and speech applications as a way of inventorising Sesotho digital language resources.

2 Background

Natural Language Processing (NLP) has been a very active area of research (Drake 2003). It is aimed at both automatically analysing and representing human language using computational technologies (Cambria & White 2014, Kang et al. 2020). These computational techniques are based on both theory and available technologies that enable them to learn, from existing human language content, understand how the human language is produced, and produce human language content independent of humans (Liddy 2001, Hirschberg & Manning 2015). NLP research gathers knowledge on how humans use language and in turn develop applications to enable computers to simulate and handle natural languages (Chowdhury 2003). As such, NLP is a discipline within Artificial Intelligence (AI) and linguistics since it is characterized by human-like language processing capabilities (Drake 2003, Nadkarni et al. 2011). Frequent applications of NLP include (i) information retrieval, (ii) information extraction, (iii) question answering, (iv) summarization, (v) machine translation, and (vi) dialogue systems (Drake 2003). Some models cater for more than one language (Castellucci et al. 2021).

Carrying out NLP tasks relies on the availability of HLT resources which are developed using digital language resources. Unfortunately, none of the South African official languages have official Basic Language Resource Kits (BLARKs) as conceptualised in Krauwer (2003). Generally, a BLARK defines the minimal required resources for performing pre-competitive research in both spoken and written language for a specific language (Arppe et al. 2010). BLARKs are categorised into three, namely, definition, specification, and content (Maegaard et al. 2006).

First, the definition category indicates what should be regarded as basic in the given language. For instance, Arppe et al. (2016) present a matrix that indicates the resources needed for Plains Cree (an indigenous language of central Canada) together with their levels of importance.

Second, the specification category prescribes the quantities of the resources defined such as the Arabic basic resource specification by Maegaard et al. (2006) that lists the quantities of resources needed for Arabic. Finally, the content category describes what already exists. One example of this is a survey of applications available for Swedish (Elenius et al. 2008).

In this article, we limit our discussion of basic digital language resources to BLARK content. That is, we inventorize written and spoken digital language resources available in and for Sesotho. Although the definition of BLARK content is language-independent, the resources identified and the level of importance allotted to the resources are specific to Sesotho (Krauwer 2003, Arppe et al. 2010).

Amongst others, industry and education institutions benefit from the availability of BLARKs (Maegaard et al. 2005). In other words, since the BLARK can identify both what already exists and what should be developed, industrial and academic developers can earmark needed resources. Furthermore, BLARK content makes it easier for researchers and educators to unearth paid and freely accessible resources at their disposal. Even so, there are two main issues with BLARKS. First, availability is not a binary distinction in that some existing resources may still be inaccessible due to financial or copyright issues (Krauwer 2003). Second, the multiplicity of resources may not be equated to usability,

quality output, or user-friendliness.

When carrying out BLARK content, three BLARK aspects are evaluated, namely, (i) data, (ii) modules, and (iii) applications (Arppe et al. 2010, Krauwer 2003). For instance, in their discussion, Sibeko & Van Zaanen (2022*a*) identify their target application, that is, a readability analysis application. However, they indicated that some modules such as a syllabification system needed to be developed. For the development of the syllabification system, they created a corpus of syllabification annotated corpora. In this way, the data enabled the development of the module which will then be incorporated into their application.

3 Methodology

It is recommended that the South African Centre for Digital Language Resources (SADiLaR) purposefully manages all publications of language resources in South African languages (Moors, Wilken, Gumede & Calteaux 2018). SADiLaR is supported and funded by the South African National Department of Science and Innovation (Sefara et al. 2021). SADiLaR supports development and innovation in the official languages of South Africa. According to Wilken et al. (2018), SADiLaR also aims to facilitate access to digital data and software applications. To this end, it has a publicly accessible repository [https://repo.sadilar.org].

The BLARK content presented in this article is based on digital language resources indexed in the afore mentioned repository. In the first stage of our research, we used the search functionality on SADiLaR's website to extract the indexed language resources. A very broad search query: "Sesotho" was used to yield a total of 123 resources. A summary of these results is presented in Table 1. In the second stage, we searched google scholar for literature related to the digital language resources identified from SADiLaR's repository. Our analysis includes both general language resources that were developed for other languages and can be applied to Sesotho, and resources that were specifically developed for Sesotho. We do this since overviews of this kind ought not be limited to monolingual resources (Arppe et al. 2010). Search results were independently evaluated by the researchers and the results were comparatively analysed.

4 Findings

As indicated above, we limit our discussion to BLARK content as defined by Maegaard et al. (2006). We discuss written resources, speech-based resources, modules, and applications available to and for Sesotho.

4.1 Written Resources

4.1.1 Monolingual lexicon

Few written monolingual corpora were identified. One, the National Centre for Human Language Technology (NCHLT) produced four data sets focused on plain corpora, annotated corpora (Eiselen & Puttkamer 2014), phrase chunks (Eiselen 2016*b*), and named entity (Eiselen 2016*a*). Two, a customised government domain specific dictionary was identified (Bosch & Griesel 2017). Furthermore, a genre classification corpus for Afrikaans, Sepedi, Setswana, isiXhosa, isiZulu, and Sesotho was also identified. The corpus is composed of poetry, advertisements, informational pamphlets, instructions, news, official texts like policies, and speech texts (Snyman et al. 2011). Finally, a syllable annotated word list was identified. It contains one word and its corresponding syllabified versions on each line (Sibeko & Van Zaanen 2022*b*). The word list contains a set of 1355 entries extracted from an existing bilingual Sesotho-English dictionary (Chitja 2010).

Table 1: Search Query Results

Results	Sum
Total results	123
Modules and applications	48
Spoken corpora	16
Written corpora	19
Relevant results	83

Table 2: Translated word lists available in Sesotho

word list	Size
Election	559
Parliamentary jargon	502
HIV/AIDS	586
Arts and Culture for the intermediate phase	550
Mathematics	984
Natural Sciences and Technology list	2756
Information and Communication Technology	132
Life Orientation for the intermediate phase	1628
Soccer	297
Gender Terminology List	446

4.1.2 Translated word lists

Many language pairs lack enough parallel texts (Koehn & Knight 2002). This is seemingly the same case in South African indigenous languages. Even so, there are few translations that use English as a pivot language. We identified a total of nine such word lists that are translated from English to the other official languages of South Africa. The word lists are presented in Table 2.

The word lists were commissioned by the National Language Services under South Africa's Department of Sport, Arts and Culture. The word lists present singular word translations on different subject matters such as politics, education and sports. Language and translation experts in different official languages of South Africa gather in a workshop setting and words are translated in groups. Group quality assurance workshops are then held to ensure quality translations.

Two written dictionary data sets were identified. One, the Bukantswe Sesotho-English bilingual dictionary word list contains a total of 10085 bilingual segments in Sesotho and English. Each line presents the Sesotho term followed by the English translation and the relevant part of speech where available. Two, the Sesotho custom dictionary for government domain is a word list that contains two types of words, namely, those that are exclusive to the government domain and those that do not follow official Sesotho orthographic conventions.

We also identified the Mburisano Covid-19 multilingual corpus that contains screening and triage vocabulary of Covid-19 related multilingual corpora in all official languages. The corpus was developed in response to the Covid-19 pandemic as an attempt to ensure access to information for people of different linguistic repertoires. English is used as the source text for all the other official languages. Unfortunately, the corpus is not clearly marked for specific languages. Although this resource is called a corpus, it contains a translated word list.

4.1.3 Machine Translated corpora

We identified three Machine Translation (MT) corpora. Two of the corpora were produced as part of the Autshumato MT Translation Memory (TM) project. Both the multilingual word and phrase translations and the MT evaluation set contain aligned translations from English to the other ten official languages. The third MT corpus, the Centre for Text Technology (CText) multilingual text corpora, provides document level aligned texts for MT purposes.

4.2 Spoken resources

We initially identified a total of 18 speech related resources. However, upon close scrutiny, we found that two of these resources were false results. One such instance is the Lwazi II Cross-lingual Proper Name corpus that is meant for Northern Sotho and not Sesotho (Kgampe & Davel 2010, 2011). In the end, we discuss a total of 16 spoken language resources.

4.2.1 Text to Speech and Automatic Speech Recognition

Seven resources were identified in this category. Three sets of corpora were produced by the Lwazi project. Two of the corpora are purposed for text-to-speech (TTS) while one is aimed at automatic speech recognition (ASR). The Lwazi and Lwazi II Sesotho TTS corpora contain transcriptions annotated with phonemic and orthographic informa-

tion. The training sentences contain approximately equal speech sounds (Badenhorst et al. 2011). The Lwazi Sesotho ASR corpus contains audios and transcriptions used for the Lwazi speech recognition systems. The transcriptions are annotated with orthographic information for each word.

The NCHLT project also produced three TTS corpora. The speech corpus contains orthographically transcribed broadband speech data including a text suite of eight speakers (De Vries et al. 2014). We also identified the auxiliary speech corpus (Barnard et al. 2014, De Vries et al. 2014, Badenhorst et al. 2019). Finally, the inlang Pronunciation Dictionary for Sesotho contains an associated rule for generating pronunciations for unseen words (Barnard et al. 2014, Davel et al. 2013).

Lastly, we identified the Sesotho multi-speaker TTS corpus that contains audio and annotated transcriptions created for investigating the implementation of a high-quality TTS system that uses a low-cost process. The data sets were quality checked. However, the read.me file indicates that there might still be some errors. Unfortunately, accuracy results were not reported.

4.2.2 Sound-based corpora

In this category, we discuss tone-based and pronunciation-based corpora. Four tone-based corpora were identified. First, the Sesotho vowel speech data set contains a collection of words that represent five Sesotho orthographic vowels, that is, —*a e i o u*. The speech data was recorded with seven females and three male participants. Second, the metadata indicates that the intonation model for Bantu tone languages contains a model for isiZulu, Sepedi, Setswana, and Sesotho. The model is intended for theoretical linguistic intonation rules in prose. Unfortunately, the model has not been uploaded on the repository. As such, we are unable to report on its contents. Third, the Sesotho tone data set also contains male and female audio recordings. The participants were sourced from a specific region of South Africa called Qwaqwa. Fourth, the Sesotho function word speech data corpora, contains audios and annotated transcriptions aimed at studying the role of tone in *ke* and *o* as function words.

One pronunciation-based corpus was identified. The South African Multilingual Proper Names Corpus (Multipron) was developed in response to accent based variation in the pronunciation of personal names (Giwa et al. 2011). This corpus uses different speakers from isiZulu, Afrikaans, English and Sesotho. Four participants read Sesotho words. Sesotho words comprise 15% of the total data collected in this corpus. Each directory consists of the orthographic transcription in a text file, phonemic transcription containing phoneme strings and an audio file consisting of an acoustic representation of each word.

4.2.3 Dictation and telephony data

We identified a total of three digital language resources in this category. First, the South African Directory Enquiries (SADE) Name corpus (Thirion et al. 2020), uses a Sesotho home language speaker for accent, but the words used for training the voice dictation platform do not contain Sesotho names. Second, we identified two telephony speech data sets. The High quality TTS data for Afrikaans, Setswana, isiXhosa and Sesotho contains multi-speaker TTS audio data and transcription files. The African Speech Technology Sesotho speech corpus contains speech spoken by Sesotho mother tongue speakers (Roux et al. 2004). Third, the SADE municipality hotline IVR prompts corpus contains audio and corresponding transcriptions in English, isiZulu and Sesotho (Van Heerden et al. 2014). The produced recognition system recognises pronunciations with Afrikaans, English, isiZulu and Sesotho accents. The interface can also be customised to isiZulu or Sesotho.

4.3 Modules and Applications

This section discusses a total of 48 modules and applications available to Sesotho. Some applications have older and newer versions while others have ba-

sic and professional versions.

4.3.1 Non-language specific applications

Six non-language specific applications were noted. First, the Autshumato PDF Text Extractor is used for extracting texts for translation using the Autshumato automatic translation machine. It functions as a plugin for the OmegaT computer-assisted translation application. The translation system is named after Autshumato, possibly South Africa's first official translator and interpreter (Groenewald & Fourie 2009, Skosana & Mlambo 2021). Autshumato is one of South African government's initiatives for improving multilingualism through an increase in both quantity and quality of translation services (Groenewald & Fourie 2009). Unfortunately, the quality and accuracy of the machine translation web service is only optimal for government data because it was trained on this type of data (Skosana & Mlambo 2021). Nonetheless, if a translator translates similar documents, they can save the translation memories (TMs) and rely on them.

Second, we identified the Autshumato translation memory exchange (TMX) integrator which works as a utility that enables merging multiple TMs over networks through subversion (Schlemmer & Fourie 2013). Unfortunately, translation memory sharing is not yet common practice. To this end, we cannot estimate the possible contribution that individual translation projects can make towards building a bigger and more reliable translation memory for Sesotho translations. The TMX integrator only supports translation from English to Sesotho and not from Sesotho to English (Reina et al. 2013).

Third, the DictionaryMaker (Davel & Barnard 2003), has been evaluated on German, but we hope that it can also be applied on Sesotho texts. The DictionaryMaker allows the user to develop a pronunciation dictionary. When used, the human effort needed for developing such a dictionary is decreased. Moors, Wilken, Gumede & Calteaux (2018) indicate that there are three pronunciation dictionaries for Sesotho, one identified in 2009 and two identified in the 2014 audit, namely, the NCHLT-inlang, Lwazi and Lwazi II pronunciation dictionaries.

Fourth, we identified corpus related applications such as the (i) CorpusCatcher, designed to crawl the web for data using seed documents for constructing queries for document retrieval, (ii) Spelt, used in the creation of classified word lists that are used in spell checking, and (iii) TurboAnnotate1.0, used for manual creation of gold standard and annotated lists. According to Van Huyssteen & Puttkammer (2007), this application lowers human effort and improves accuracy of annotation. The TurboAnnotate1.0 application uses the Tilburg Memory-Based Learner machine learning system (*see* Daelemans et al. (2004)). It allows mother tongue speakers with limited and no experience with computational linguistics to annotate texts. Machine learning then learns from the user generated annotations.

4.3.2 South African official languages

In this section, we discuss applications that were developed for all eleven South African official languages. We identified at least 30 applications in this category.

Five NCHLT products were identified, namely, the Optical Character Recognition (OCR), language identifier, text web service, tagger and Part of Speech tagger. The OCR for South African Languages (Hocking & Puttkammer 2016) enables the user to convert scanned documents into editable texts. It can reproduce almost any character or image. The Language Identifier uses both a graphical user interface and a command line interface for automatically identifying official South African languages. The text web services provides access to tokenisers, sentence separators, POS taggers, phrase chunkers, named entity recognisers and OCR in South African languages. The tagger can be used either through the command line or a user interface. It annotates running text with either POS, named entity,

noun phrase chunks, or nouns. Finally, the Part of Speech Taggers were developed using a minimum of one million government published tokens per language (Eiselen & Puttkamer 2014).

Four CText products were identified. First, the Alignment Interface and the Alignment Interface Pro are utility applications used for aligning source texts. The Pro version allows for editing the segments. These products work with the CText applications 1 that allows for automatic corpus query and manipulation for tokenisation and sentencisation, frequency and word list extraction, searching and extracting collocations. Additionally, the CText Applications 2 adds POS tagging, named entity recognition, and phrase chunking.

We also identified six independent applications, namely, (i) the AStudio, a software that incorporates a graphic interface for the developing flowcharts for speech applications, (ii) Automatic Oral Proficiency Assessment application, developed as part of the Development of Resources for Intelligent Computer-Assisted Language Learning project, (iii) the Language Identifier (LID) classifier token level classification for all official languages, (iv) a Combination Tagger that uses memory-based tagger (MBT), support vector machines (SVM), Mobotix part of speech tagger (MXPOST) and Trigrams'n'Tags (TnT) for deciding on tags, (v) the South African Fonts collection that contains fonts representing all alphabets and characters used in South African official languages, and (vi) the Format Normaliser 1.0. for normalising input files to utf8 txt, replacing smart quotes, and removing empty lines.

Four translation resources were also identified. One, the Rhonda machine translation system can handle speech to speech translation, or speech to text translations. Two, the Translate application kit 1.4.0 is a collection of applications and parsers that handles various localisable and translatable formats. It composes modules for segmentation, authentication and text enumeration. Three, the Autshumato Translation Management Systems web applications allow for capturing, editing, exporting and importing terminologies. Four, the Autshumato Text Anonymiser classifies and replaces sensitive information. For instance, if one wishes to anonymise sensitive information such as study participants' names, this application replaces them with pseudo names.

Five TTS related applications were also identified in this category. One, the Lwazi Telephony Platform combines Asterisk with MobillVR Python interface in one unified control interface. In this way, the process of developing experimental applications is accelerated. Two, the Qfrency TTS phone mappings application maps is used with Lwazi and NCHLT pronunciation dictionaries of the official languages. Three, the multilingual TTS Speect system provides a full service of decoding and encoding texts, that is, text analysis and speech synthesis with various APIs. Furthermore, it can be used for research and development of TTS system voices. Four, the Phonetic aligner contains scripts for automatic phonetic alignment of speech corpora using the hidden markov models. Finally, the Text Selection scripts for ASR/TTS, uses phonetic rules to phonetise texts, then the diphones are used for TTS and triphones are used for ASR on a per language basis.

4.3.3 Language specific resources

We identified four language specific applications that also incorporate Sesotho. One, the EtsaTrans translation system user interface is available in English, Afrikaans, isiXhosa and Sesotho. Two, the Multilingual Illustrated Dictionary with interactive games application is available for seven of official languages, namely, Afrikaans, South African English, isiXhosa, isiZulu, Sepedi, Setswana and Sesotho. Three, the NWU TransTips 1.0 is a PhP programming script that browses the web page for terms in the database. The translations appear when the user hovers over a certain word. Unfortunately, the user still has to decide on the correct translation between those presented.

Lastly, the Automated multilingual telephone access to financial services is a prototype that allows

switching between isiZulu, isiXhosa, English and Sesotho.

4.3.4 Sesotho-only

We identified seven applications that are specifically for Sesotho. First, we identify two NCHLT products. One, the lemmatiser was developed using a rules-based approach (Eiselen & Puttkamer 2014). Two, the Morphological Decomposer that splits tokens to morphemes was developed after the POS tagger (Eiselen & Puttkamer 2014). Although Eiselen & Puttkamer (2014) provide examples of decomposition for isiZulu and Afrikaans, there are no examples for decomposition in other languages such as Sesotho. Second, the Lwazi Sesotho Pronunciation Dictionary includes audios for phonemes and letter-to-sound rule set based on generic words (Davel & Martirosian 2009). Although accuracy is not guaranteed, practical usability is assured. Third, the Lwazi II Sotho Pronunciation Dictionaries (Du Plessis et al. 1974), are based on the Lwazi dictionaries. Fourth, the Spelling Checker 1.0 is an application that checks spelling and provides hyphenations. It is compatible with some versions of Ms Office. CText continues to improve this application.

Finally, syllabification systems were developed as part of a project on developing a metric for measuring text readability in Sesotho. Two systems are part of the package. One, the ML-Based system produces 78.97% accurate results. Two, the rules-based system produces 99.69% accurate results (Sibeko & Van Zaanen 2022*a*).

4.3.5 Games

We identified one game, the Open Spell (v1.0), a spelling game that contains spelling exercises aimed at teaching spelling skills to school children. The source code is also freely accessible.

5 Discussion and conclusion

Many LRL's are underrepresented in NLP tasks because of insufficient corpora needed to complete NLP tasks (Mahloane & Trausan-Matu 2015). The limited availability of corpora as indicated in this article shows a great need for curating more corpora for Sesotho. This article considered basic digital language resources available to Sesotho from a BLARKs content perspective.

The listed resources were listed without in depth analysis of each resource such as considering how each resource works or the levels of accuracy and practical issues such as user-friendliness. We consulted literature relevant to the resources presented in the repository. To this end, some of these resources have been judged on issues such as accuracy and user-friendliness. However, this was not the aim of this paper. This is typical of BLARK content studies. Although this is basic, it functions as a starting point for investigations of BLARKs specification and definition. That is, determining what should be regarded as basic in the language in Sesotho and the quantities that should be developed. Nonetheless, it is clear from this article that written resources are focused only on very basic functions. For instance, there are no resources for higher level basic functionalities such as semantic analysis and term extractors. Even so, only the handwritten OCR and ontologies are missing from the written resources as identified in the BLARKs (Maegaard et al. 2005, 2006, Krauwer 2003). There have been even fewer attempts at speech technologies, especially those that are specifically aimed at Sesotho or the Sesotho language group. Even so, it is interesting how much work has been achieved.

A narrowed investigation into the current resources and an evaluation of each technology should be considered for future studies. SADiLaR is currently conducting their routine language resource audit, more resources might be added to their repository after their investigation. Perhaps we may gain more insight into digital language resources available to Sesotho.

This article was limited in three aspects. First, the resources surveyed in this review are not evaluated for their usability, accuracy, and precision. Sec-

ond, some of the resources were only listed and indexed on the repository, however, the actual resources were not accessible. Third, the study only reviews resources that are indexed by SADiLaR. We recommend that future developments of digital language resources for Sesotho consider this inventory in their decisions on what resources to develop so that focus is paid to new and currently unavailable resources. Even so, we acknowledge that the current index indicates variation in the current collection.

Acknowledgements

We acknowledge Professor Menno van Zaanen for his guidance in the writing of this article and for suggesting that we collaborate on this project.

References

Arppe, A., Beck, K., Branco, A., Camilleri, V., Caselli, T., Cristea, D., Hinrichs, E., Liin, K., Nissinen, M., Parra, C., Rosner, M., Schuurman, I., Skadina, I., Quochi, V., Van Uytvanck, D. & Vogel, I. (2010), Description of the BLARK, the situation of individual languages, Report, Clarin.

Arppe, A., Lachler, J., Trosterud, T., Antonsen, L. & Moshagen, S. N. (2016), Basic language resource kits for endangered languages: A case study of plains cree, *in* 'Proceedings of the 2nd Workshop on Collaboration and Computing for Under-Resourced Languages Workshop (CCURL 2016)', pp. 1–8.

Badenhorst, J., Martinus, L. & De Wet, F. (2019), Blstm harvesting of auxiliary nchlt speech data, *in* 'Proceedings of South African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/ROBMECH/PRASA 2019)', pp. 123–128.

Badenhorst, J., Van Heerden, C., Davel, M. & Barnard, E. (2011), 'Collecting and evaluating speech recognition corpora for 11 South African languages', *Language resources and evaluation* pp. 289–309.

Barnard, E., Davel, M. H., Van Heerden, C., F. De Wet, F. & Badenhorst, J. (2014), The nchlt corpus of the South African languages, *in* 'Proceedings of the 4th International Workshop Spoken Language Technologies for Under-resourced Languages', pp. 194–200.

Bosch, S. & Griesel, M. (2017), 'Strategies for building wordnets for under-resourced languages: the case of African languages', *Literator* .

Cambria, E. & White, B. (2014), 'Jumping NLP curves: A review of natural language processing research', *IEEE Computational intelligence magazine* **9**, 48–57.

Castellucci, G., Filice, S., Croce, D. & Basili, R. (2021), Learning to solve NLP tasks in an incremental number of languages, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers)', pp. 837–847.

Chiguvare, P. & Cleghorn, C. W. (2021), Improving transformer model translation for low resource South African languages using bert, *in* 'IEEE Symposium Series on Computational Intelligence (SSCI)', IEEE, pp. 1–8.

Chitja, M. (2010), *Phatlamantsoe ya Sesotho ya Machaba*, Mazenod Publishers.

Chowdhury, G. (2003), 'Natural language processing', *Annual Review of Information Science and Technology* **37**, 51–89.

Cieri, C., Maxwell, M., Strassel, S. & Tracey, J. (2016), Selection criteria for low resource language programs, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 4543–4549.

Daelemans, W., Zavrel, J., Van Der Sloot, K. & Van den Bosch, A. (2004), Timbl: Tilburg memory-based learner, Tehnical report, Tilburg University.

Davel, M. & Barnard, E. (2003), Bootstrapping in language resource generation, *in* 'Fourteenth Annual Symposium of the Pattern Recognition Association of South Africa', pp. 97–100.

Davel, M., Basson, W., Charl, V. H. & Barnard, E. (2013), 'Nchlt dictionaries: Project report'.
URL: `https://sites.google.com/site/nchltspeechcorpus/home`

Davel, M. & Martirosian, O. (2009), Pronunciation dictionary development in resource-scarce environments, *in* 'Proceedings of the Interspeech', pp. 2851–2854.

De Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., De Wet, F., Barnard, E. & De Waal, A. (2014), 'A smartphone-based ASR data collection tool for under-resourced languages', *Speech Communication* pp. 119–131.

Drake, M. (2003), *Encyclopedia of Library and Information Science, Second Edition*, Taylor & Francis.

Du Plessis, J. A., Gildenhuys, J. G. & Moiloa, J. J. (1974), *Moiloa. Bukantswe ya malemepedi Sesotho-Seafrikanse / Tweetalige woordeboek Afrikaans-Suid-Sotho*, first edition edn, Via Afrika Beperk.

Eiselen, E. R. & Puttkamer, M. J. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the 9th International Conference on Language Resources and Evaluation', pp. 3698–3703.

Eiselen, R. (2016*a*), Government domain named entity recognition for South African languages, *in* 'Proceedings of the 10th Language Resource and Evaluation Conference'.

Eiselen, R. (2016*b*), South African language resources: phrase chunkers, *in* 'Proceedings of the 10th Language Resource and Evaluation Conference'.

Elenius, K., Forsborm, E. & Megyesi, B. (2008), Language resources and tools for swedish: A survey, *in* 'Proceedings of the LREC 2008'.

Giwa, O., Davel, M. H. & Barnard, E. (2011), A Southern African corpus for multilingual name pronunciation, *in* 'Pattern Recognition Association of South Africa and Mechatronics International Conference'.

Groenewald, H. J. & Fourie, W. (2009), Introducing the autshumato integrated translation environment, *in* 'Proceedings of the 13th Annual conference of the European Association for Machine Translation', pp. 190–196.

Grover, A. S., Van Huyssteen, G. B. & Pretorius, M. W. (2010), The South African human language technologies audit, *in* 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', pp. 2847–2850.

Grover, A. S., Van Huyssteen, G. B. & Pretorius, M. W. (2011), 'The South African human language technology audit', *Language resources and evaluation* **45**, 271–288.

Hanslo, R. (2021), Evaluation of neural network transformer models for named-entity recognition on low-resourced languages, *in* '16th Conference on Computer Science and Intelligence Systems (FedCSIS)', IEEE, pp. 115–119.

Hirschberg, J. & Manning, C. D. (2015), 'Advances in natural language processing', *Science* **349**, 261–266.

Hocking, J. & Puttkammer, M. (2016), Optical character recognition for South African languages, *in* 'Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)', pp. 1–5.

Kadenge, M. & Mugari, V. (2015), 'The current politics of African languages in zimbabwe', *Per Linguam: a Journal of Language Learning* **31**, 21–34.

Kang, Y., Cai, Z., Tan, C. W., Huang, Q. & Liu, H. (2020), 'Natural language processing (NLP) in management research: A literature review', *Journal of Management Analytics* **7**, 139–172.

Kgampe, M. & Davel, M. H. (2010), Consistency of cross-lingual pronunciation of south-African personal names, *in* 'Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)', pp. 123–127.

Kgampe, M. & Davel, M. H. (2011), The predictability of name pronunciation errors in four South African languages, *in* 'Proceedings of the Pattern Recognition Association of South Africa annual symposium (PRASA)', pp. 85–90.

Koai, M. & Fredericks, B. G. (2019), 'Sesotho is still a marginalised language', *Southern African Linguistics and Applied Language Studies* **37**, 303–314.

Koehn, P. & Knight, K. (2002), Learning a translation lexicon from monolingual corpora, *in* 'ACL Special Interest Group on the Lexicon (SIGLEX)', pp. 9–16.

Krauwer, S. (2003), The basic language resource kit (BLARK) as the first milestone for the language resources roadmap, *in* 'Proceedings of SPECOM', Vol. 2003, pp. 15–22.

Liddy, E. D. (2001), *Natural Language Processing*, 2nd ed edn, Marcel Decker, Inc.

Maegaard, B., Choukri, K., Mokbel, C. & Yaseen, M. (2005), *Language technology for Arabic*, Center for Sprogteknologi, University of Copenhagen.

Maegaard, B., Krauwer, S., Choukri, K. & Jørgensen, L. D. (2006), The BLARK concept and BLARK for arabic, *in* 'LREC', pp. 773–778.

Magueresse, A., Carles, V. & Heetderks, E. (2020), 'Low-resource languages: A review of past work and future challenges', *arXiv:2006.07264* .

Mahloane, M. J. & Trausan-Matu, S. (2015), Metaphor annotation in Sesotho text corpus: towards the representation of resource-scarce languages in NLP, *in* '20th International Conference on Control Systems and Computer Science', IEEE, pp. 405–410.

Moeketsi, V. S. M. (2014), 'The demise of Sesotho language in the democratic South Africa and its impact on the socio-cultural development of the speakers', *Journal of Sociology and Social Anthropology* **5**, 217–224.

Mojela, V. (2016), Etymology & figurative: The role of etymology in the lemmatization of Sotho terminology, *in* 'The 10th International Conference of the Asian Association for Lexicography (AsiaLex2016)', IEEE, pp. 93–100.

Moors, C., Wilken, I., Calteaux, K. & Gumede, T. (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', pp. 296–304.

Moors, C., Wilken, I., Gumede, T. & Calteaux, K. (2018), 'Human Language Technology Audit 2017/18'.
URL: `https://sadilar.org/index.php/en/2-general/284-health-resources`

Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. (2011), 'Natural language processing: an introduction', *Journal of the American Medical Informatics Association* **18**, 554–551.

Ndlovu, E. (2011), 'Mother tongue education in the official minority languages in zimbabwe', *South African Journal of African Languages* **31**, 229–242.

Ndlovu, E. (2013), Mother tongue education in official minority languages of Zimbabwe: A language management critique, Thesis, University of the Free State.

Nkolola-Wakumelol, M., Rantsoz, L. & Matlhaku, K. (2012), Syllabification of consonants in Sesotho and Setswana, *in* H. S. Nginga-Kouma-Binza & S. Bosch, eds, 'Language Science and Language technology in Africa: Festschrift for Justus C. Roux', Sun Express, Stellenbosch, South Africa, pp. 10–13.

Pretorius, L., Soria, C. & Baroni, P., eds (2014), *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, LREC.

Reina, A., Robles, G. & González-Barahona, J. M. (2013), A preliminary analysis of localization in free software: how translations are performed, *in* 'IFIP International Conference on Open Source Systems', pp. 153–167.

Riep, D. M. (2013), 'Seeing Sesotho: art, history, and the visual language of South Sotho identity', *Southern African Humanities* **25**, 217–244.

Roux, J. C., Louw, P. H. & Niesler, T. R. (2004), The African speech technology project: An assessment, *in* 'Proceedings of LREC', pp. 93–96.

Schlemmer, M. & Fourie, W. (2013), 'Autshumato tmx integrator'.
URL: `https://repo.sadilar.org/handle/20.500.12185/416`

Sefara, T. J., Mokgonyane, T. B. & Marivate, V. (2021), Practical approach on implementation of wordnets for South African languages, *in* 'Proceedings of the 11th Global Wordnet Conference', Global WordNet Association, pp. 20–25.

Sibeko, J. & Van Zaanen, M. (2022*a*), Developing a text readability system for Sesotho based on classical readability metrics, *in* 'Proceedings of Digital Humanities Conference: Responding to Asian diversity', Vol. 2022.

Sibeko, J. & Van Zaanen, M. (2022*b*), 'Raw and syllabified word list for Sesotho'.
URL: `https://repo.sadilar.org/handle/20.500.12185/556`

Skosana, N. J. & Mlambo, R. (2021), 'A brief study of the autshumato machine translation web service for South African languages', *Literator* pp. 1–7.

Snyman, D., Van Huyssteen, G. B. & Daelemans, W. (2011), Automatic genre classification for resource scarce languages, *in* 'Proceedings of the 23rd Annual Symposium of the Pattern Recognition Association of South Africa', pp. 132–137.

Strassel, S. & Tracey, J. (2016), Lorelei language packs: Data, tools, and resources for technology development in low resource languages, *in* 'Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)', pp. 3273–3280.

Thirion, J. W., Van Heerden, C., Giwa, O. & Davel, M. H. (2020), 'The South African directory enquiries (sade) name corpus', *Language Resources and Evaluation* pp. 155–184.

Van Heerden, C., Kleynhans, N., Barnard, E. & Davel, M. (2010), Pooling ASR data for closely related languages, *in* L. Besacier & E. Castelli, eds, 'Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages (SLTU 2010)', School of Computer Sciences, Universiti Sains Malaysia, pp. 17–23.

Van Heerden, E., Davel, M. & Barnard, E. (2014), Performance analysis of a multilingual directory enquiries application, *in* 'Proc. Annual Symp. Pattern Recognition Association of South Africa (PRASA)', pp. 258–263.

Van Huyssteen, G. B. & Puttkammer, M. (2007), 'Accelerating the annotation of lexical data for less-resourced languages', *Interspeech* pp. 1505–1508.

Wilken, I., Gumede, T., Moors, C. & Calteaux, K. (2018), Human Language Technology Audit 2018: Design considerations and methodology, *in* 'International Conference on Intelligent and Innovative Computing Applications (ICONIC)', IEEE, pp. 1–7.

Wissing, D. & Roux, J. C. (2017), 'The status of tone in Sesotho: a production and perception study', *Nordic Journal of African Studies* **26**, 19–19.

Creating electronic resources for African languages through digitisation: a technical report

Taljard, Elsabé
University of Pretoria
elsabe.taljard@up.ac.za

Prinsloo, Danie
University of Pretoria
danie.prinsloo@up.ac.za

Goosen, Michelle
University of Pretoria
michelle.goosen@up.ac.za

Abstract

The need for electronic resources for (under-resourced) African languages is an often stated one. These resources are needed for language research in general, and more specifically for the development of Human Language Technology (HLT) applications such as machine translation, speech recognition, electronic dictionaries, spelling and grammar checkers, and optical character recognition. These technologies rely on large quantities of high-quality electronic data. Digitisation is one of the strategies that can be used to collect such data. For the purpose of this paper, digitisation is understood as the conversion of analogue text, audio and video data into digital form, as well as the provision of born digital data that is currently not available in a format that enables downstream processing. There is a general perception that the African languages are under-resourced with regard to sufficient digitisation tools to function effectively in the modern digital world.

Our paper is presented as a technical report, detailing the tools, procedures, best practices and standards that are utilised by the UP digitisation node to digitise text, audio and audio-visual material for the African languages. The digitisation effort is part of the South African Digital Languages Resources (SADiLaR) project (https://www.sadilar.org/index.php/en/), funded by the Department of Science and Innovation. Our report is based on a best practices document, developed through the course of our digitisation project and forms part of the deliverables as per contractual agreement between the UP digitisation node and the SADiLaR Hub. The workflow as explained in this document was designed with this specific project in mind; software and hardware utilised were also selected based on the constraints with regard to capacity and available technical skills in mind. We motivate our choice of Optical Character Recognition (OCR) software by referring to an earlier experiment in which we evaluated three commercially available OCR programmes. We did not attempt a full-scale evaluation of all available OCR software, but rather focused on selecting one that renders high quality outputs. We also reflect on one of the challenges specific to our project, i.e. copyright clearance. This is particularly relevant with regard to published material. In the absence of newspapers for specifically the African Languages (isiZulu being a notable exception), the biggest portion of textual material available for digitisation consists of printed material such as textbooks, novels, dramas, short stories and other literary genres. The digitisation process is driven by the availability of material for the different languages. Furthermore, obtaining copyright clearance from publishers is a prerequisite for digitisation and especially for the release of any digitised text data for further use and / or processing. Having information on a relatively small-scale digitisation workflow and best practices readily available will enable other interested parties to participate in the digitisation effort, thus contributing to the collection of electronic data for the African languages.

Keywords: digitisation, Optical Character Recognition (OCR), copyright, metadata, electronic resoources for African Languages

1 Text digitisation

For the purpose of text digitisation, we use an Epson DS-50000 portable flatbed scanner. Camera scanners are problematic in terms of quality and consistency.

Once the resources to be digitised have been identified, the software *ABBYY FineReader 14*, an OCR application, is used for the scanning process. The decision to utilise this particular software programme is informed by a small experiment, reported on in detail in Prinsloo, Taljard and Goosen (to appear). In this experiment, two commercially available OCR programmes i.e. *ABBYY FineReader 14* and *Omnipage Professional 18,* and one locally developed scanning package, *CTexTools* were compared on good quality printouts from President Cyril Ramaphosa's 2020 state of the nation (SONA) address (https://www.gov.za/ve/speeches/president-cyril-ramaphosa-2020-state-nation-address-13-feb-2020-0000) with reference to percentage of scanning errors and overall accuracy rate. Afrikaans, isiZulu, Sepedi and Tshivenḓa were used as test languages. Our results indicated that *ABBYY* would be the preferred OCR tool for languages not utilizing more than a minimum of diacritic signs, even though those languages may not be specifically supported by the software. For Sepedi, for example, the software does not recognize the frequently occurring *š*, but activating Slovenian as the proofing language does support this character. In our experiment, the average accuracy rate for the three packages are as follows:

	ABBYY	*Omnipage*	*CTexTools*
Afrikaans	99.64	99.14	99.10
Sepedi	99.72	96.30	99.52
isiZulu	99.55	95.23	96.81
Tshivenḓa	95.61	95.50	98.30
AVERAGE	**98.63**	**96.54**	**98.43**

Accuracy of OCR scanning is affected by the quality of the source text. Defects such as distorted text lines, skewed images and noise can reduce the recognition quality of a scanned document. *ABBYY FineReader*'s automatic image pre-processing editor can remove some of the defects that may occur in a scanned document. The tools for the correction of defects include (but are not limited to) the following (https://help.abbyy.com/en-us/finereader/15/user_guide/adjustimage/):

- *Recommended pre-processing:* The software will automatically determine and apply the necessary corrections. The corrections that may be applied include noise and blur removal, colour inversion, skew correction, straightening of text lines, corrections of trapezoid distortion and cropping of image borders.
- *Split facing pages*: When scanning a book, a scanned image will usually contain two facing pages. Facing pages are split into two images.
- *Deskew images*: Corrects skewed images.
- *Straighten text lines*: Curved text lines on images are straightened.
- *Correct trapezoid distortion*: Corrects trapezoid distortions and removes image edges not containing any useful information.
- *Rotate and flip*: Images can be flipped vertically or horizontally to get them facing the right direction.
- *Crop*: The software allows the user to select a pre-set scanning area size. By cropping a document, one removes unwanted edges that do not contain any useful information.
- *Invert*: Inverts colour images. The function is useful when dealing with non-standard text colouring, such as light text on a dark background.
- *Resolution*: Changes the resolution of an image.
- *Brightness and contrast*: Changes the brightness and contrast of an image.
- *Levels*: The colour levels of the images can be adjusted by changing the intensity of shadows, light and halftones.
- *Eraser*: Erases a part or parts of images.

- *Remove colour marks*: Removes any colour stamps and marks made in pen to facilitate the OCR of the text hidden by such marks.

The default setting at which texts are scanned is 300 dots per inch (dpi). This is also the dpi recommended by *ABBYY FineReader* (https://help.abbyy.com/en-us/finereader/15/user_guide/scangeneral). In cases where the original material is poor because of age, the dpi may be increased to 600 dpi to enhance the scanning quality. As a default, scanning at 600 dpi does not seem to be feasible, since it increases the file size and the time spent on scanning, with no real improvement of the OCR scanning quality. Once a text has been scanned and the image editor used to correct defects, the scanned text is saved in an image-only PDF format. By saving the scanned document in an image-only PDF format, the document will not be searchable or contain any text layers. A second copy of the (edited) scanned document is put through the OCR function. It must be ensured that the correct language/languages are selected to enhance the OCR quality. For best results, the OCR document is first saved in UTF-8 format (a .txt file). When using ABBYY, the scanned text cannot be directly saved in Word format as the software attempts to replicate the orginal scanned document. One would also encounter scanning errors which would most likely not appear in the .txt file. Scanned documents that have been OCR'ed are then saved in PDF format. For the text cleaning process and for the running of spell checkers, the .txt files are converted to Word format.

The purpose of the cleaning process is mainly to correct scanning errors. Taking the skills level of project participants into consideration (these are mainly student assistants who have an African language as first language), we opted for a text cleaning strategy that does not need skilled computer programmers. We are aware of more sophisticated procedures such as the use of N-grams, but these require a high level of computational skill, which makes these procedures not always ideal within the context of lesser-resourced languages. We therefore rely mostly on a process of manual correction with spellchecker support. Spellcheckers are supplemented with custom dictionaries, based on word lists generated from corpora compiled by the UP digitisation node. After quality control has been carried out, the final version of the cleaned texts is stored in UTF-8 format.

In cases of born digital data, these are usually available either in .pdf or MSWord (.doc or .docx) format. In case of PDF documents, OCR scanning needs to be carried out; for Word documents, these are directly saved in UTF-8 format. Once again, it must be ensured that the PDF document should not contain a text layer. An image only PDF document allows other individuals or institutions to utilise it for the purposes of OCR research.

2 Copyright considerations

A salient aspect of text digitisation is that of copyright, especially when working with published texts such as text books, novels, dramas and other literary genres. In order to understand the complexities of copyright on digitised texts, it is important to understand the exact nature of a digitised text. In essence, digitisation is a process of converting printed texts into a machine-readable format. A digitised version involves more than a mere reproduction, as is evident from the procedure described above. As pointed out by Nicholson (2010:10), "it involves the conversion to another format, often involving modification, adaptation, or cropping, even translation, where necessary". Digitisation potentially makes information available and accessible to a wide audience and can therefore be regarded as a form of (re)publishing. Strictly speaking, the act of digitisation therefore constitutes in itself an act of infringement of copyright, unless prior clearance has been obtained from the copyright holder, which in the case of published material, is the publisher. In discussions on copyright the notion of 'fair use' is often referred to, and publishers are more

inclined to provide copyright clearance if they are convinced that digitisation amounts to fair use. 'Fair use' is determined by four factors, i.e. the purpose of the intended use, the nature of the work, the amount or substantiality used, and market impact (Besek 2003: 5; Senekal and Kotzé 2018: 267). With regard to the first factor, the distinction between commercial use and use for research purposes is relevant. Use for commercial purposes is unlikely to be viewed as fair use. Secondly, if a text is of a factual nature, rather than a creative text, the scope of fair use is generally broader. Thirdly, the smaller the portion that is digitised, the more likely it is to be regarded as fair use. It is often argued that scanning a 10% section of a text source is acceptable as fair use. However, copyright experts are quick to point out that even a single page from a text source which represent a core design can be judged as copyright infringement. Determining the effect of making a digital copy available on the potential market for the digitised text or work, constitutes the fourth factor. As Besek (2003: op cit.) points out, use that supplants the market for the original is unlikely to qualify as fair. However, deciding on whether use is 'fair' seems in many cases a subjective decision and needs to be determined on a case by case basis.

In our opinion there are no safe generic copyright rules for text scanning except for explicit permission of the copyright owner. Obtaining copyright can be simplified by negotiating the exact intended use of the data. So, for example, publishers might not agree to the digitisation of full texts as they fear that such data could be resold and consequently will lead to loss of income for the owner. Publishers might be more inclined to give permission to the use of data for research purposes or if the data will only be stored in scrambled format. Texts can, for instance, be scrambled on paragraph or sentence level which simply means that sentences and paragraphs no longer appear in the same order as the source texts. In order to safeguard the person(s) and / or institution(s) responsible for digitisation, a written contract stating the exact sources and the permitted utilization of the texts is the only option.

3 Digitisation of audio material / cassettes

The hardware used for converting audio material is USB Cassette Capture (tape to MP3 converter). Determining the quality of audio cassettes is the first step in the digitisation of audio material. Incorrect storage, deterioration because of age and physical damage to cassettes can all affect the quality of the digitised version. In cases where the quality of the original recording is less than perfect, a decision as to the usefulness of digitisation should be taken, based on the inherent value of the resource.

Audacity (https://www.audacityteam.org/about/), released 20 years ago, is open source software and is regarded as being as effective as many premium paid-for applications (https://www.techradar.com/reviews/audacity). For the purpose of the digitisation of audio cassettes, the software allows the user to digitise recordings from other media, edit the digitised file, i.e. cut, copy, paste and delete and export the digitised file to the desired format. It also has a noise reduction function that can reduce constant background sounds.

Prior to digitisation it must be ascertained whether any analogue noise reduction technique was applied to the tape when it was encoded, and the corresponding decoding filter, either in the analogue or digital domain, must be applied in the digitised copy of the analogue cassette. Examples of noise reduction systems include Dolby B and Dolby C.

The default quality settings at which an audio cassette is being digitised through Audacity, is set at 44100 Hertz (Hz) and 32-bit format in stereo. The final format in which the digitised audio cassettes is stored in is Waveform Audio File Format (WAV). Before the digitised files (.aup files) are stored in said format, the .aup files must be checked to ensure that the default settings were used for digitisation. For any .aup files not digitised according to the default settings, the

process must be repeated. Any static at the beginning and/or end of a digitised file must be removed.

4 Digitisation of video material/ cassettes

The first challenge posed by the digitisation of video material is finding VHS (Video Home Systems) video players needed for playing of video cassettes. Since this is old technology, these players are not readily available. Spare parts, such as drive belts are also only available outside of South Africa. The software used is Elgato video capture (https://www.elgato.com/en/video-capture) and is regarded as one of the best video capture devices (https://www.msn.com/en-us/Lifestyle/rf-buying-guides/best-video-capture-devices-reviews). The resolution at which videos are digitised, is 720x576p (720 pixels across and 576 pixels tall). The recommended video bitrate is 5 Megabits per second (Mbps) and the frame rate 25 frames per second (fps). The recommended audio bitrate is 224 Kilobits per second (Kbps) and the audio sample rate is 44 100 Hertz (Hz). The colour mode is Red, Green and Blue (RGB) colour space. The recording format is Digital Video Disc or Digital Versatile Disc (DVD), the recording video type is Phase Alternating Line (PAL) and the quality is set at best. The digitised version is stored in .mpg (MPEG2) format (codec: MPEG-2 video (mpgv) / codec: MPEG audio layer 2 (mpga)). MPEG, which stands for *Moving Pictures Expert Group,* is a standard audio and video coding compression. The On Screen Display (OSD) messages must be turned off and may not appear in any digitised video. It also needs to be verified that the digitisation process accounts for the analogue Dolby B encoding applied in the VHS recording standard.

5 Provision of metadata

The provision of metadata for any digitised resource is an indispensable part of the digitisation process. Burnard (2004) describes metadata as "the kind of data that is needed to describe a digital resource in sufficient detail and with sufficient accuracy for some agent to determine whether or not that digital resource is of relevance to a particular enquiry". With regard to digitised texts, metadata should ideally be presented in an integrated form, together with the text file, using the same encoding principles or markup language used in the text file itself. According to the Federal Agencies Digital Guidelines Initiative (FAGDI) (https://www.digitizationguidelines.gov/), presenting the metadata in this format facilitates the identification, management, access, use and preservation of a digital resource. It helps to ensure that the text and the metadata are kept together and can be distributed as a single unit. The TEI (Text Encoding Initiative) has been a major influence in this regard, publishing an extensive set of Guidelines for the Encoding of Machine Readable Data (TEI P1) (https://tei-c.org/release/doc/tei-p5-doc/en/html/index.html). One of the recommendations of the TEI was the definition of a specific metadata component, called the *TEI Header*. This header functions as a kind of electronic title page, providing information such as *inter alia* a file description, indication of text derivation and a bibliographic description. Once again, presenting the metadata for each digitised text in this format requires considerable computational expertise and we opted for a simpler, albeit it perhaps old-fashioned approach. We provide mostly standard bibliographic description in a separate document, providing the following information: title, name of author(s), date of publication, ISBN, publisher, genre, description of the genre, language (using ISO language codes), status of copyright, number of pages (PDF document), tokens, media type, encoding, format / file extension and name of document. The name of a text document must contain the following information: language(s), title of document, author's surname and genre, for example *zul_Zibukhipha zibuthela_Shabangu_novel*. In cases where a text document's source is a newspaper or a magazine, the title of such documents in a dataset must contain the following information:

language(s), title of document, publisher and the date, for example *sot_Boleng ba popeho tse fapaneng tsa mebu_Pula Imvula_OCT 2012*. In cases where a document contains more than one language, it should be indicated in the languages field as well as in the name of the document. The ISO abbreviations for the given languages must be written consecutively, without any spaces or delimiters, for example: *nsoengafr_Re bala Sesotho 2_Britz_reader*. The symbol used to delimit the data fields is an underscore (_).

Information included in the metadata lists of audios and videos is: title, presenter, date of publication, publisher, genre, description of genre, language (see table 1 below for ISO language codes to be used), status of copyright, length, media type, encoding, format / file extension and the name of the document. The names of audio and video files should contain the following information: language(s), title of audio/video file, presenter's surname and date, for example *afr_AFR 102 verstegniek onderrigkasset 3 kant A_Marais_19990315*. In cases where a file contains more than one language, it should be indicated in the languages field as well as in the name of the document. The ISO abbreviations for the given languages must be written consecutively, without any spaces or delimiters, for example: *zuleng_Lesson 4_presenter unknown_date unknown*. The symbol used to delimit the data fields is an underscore (_).

6 In conclusion

From the discussion above it should be clear that digitisation of especially textual material is much more than scanning a text and saving it in PDF format. In order to ensure the maximum (re-)usability of data it is extremely important that (a) data are stored in the correct format, and (b) that the correct protocols, procedures and technical guidelines are followed during the digitisation process. Apart from making digitised data available for further HLT processing and application, digitisation has an additional function, i.e. preservation of material that is invaluable and / or irreplaceable, in which case the quality of the data may be of lesser quality. As a general rule, quality of digitised material should only be compromised in the case of text, audio and video when the data is invaluable to such an extent that the user will be willing to tolerate low(er) quality for the sake of the importance of the data. This is for instance the case in very old but valuable data on audio reels or VHS video tapes damaged by moisture. In such cases a notice must be posted warning potential users of lesser quality so that users are informed that bad quality is not the result of substandard digitisation processes or equipment. Potential users can then take an informed decision as to whether they are willing to work through the data.

References

Besek, J.M. 2003. *Copyright issues relevant to the creation of a digital archive: A preliminary assessment body*. Washington, D.C.: Council on Library and Information Resources.

Burnard, L. 2004. *Metadata for corpus work*. https://www.academia.edu/3234836/Metadata_for_corpus_work. Last accessed: 25-08-2022.

Liebetrau, P. (ed.). 2010. *Managing digital collections: A collaborative initiative on the South African Framework*. Pretoria: National Research Foundation.

https://help.abbyy.com/en-us/finereader/15/user_guide/adjustimage/. Last accessed: 25-10-2022. *If your document image has defects and OCR accuracy is low.*

https://www.sadilar.org/index.php/en/. Last accessed: 12-10-2022. *South African Centre for Digital Language Resources.*

https://www.gov.za/ve/speeches/president-cyril-ramaphosa-2020-state-nation-address-13-feb-2020-0000. Last accessed: 25-10-2022. *President Cyril Ramaphosa: 2020 State of the Nation Address.*

Nicholson, D. 2010. *Copyright and related matters*. In Liebetrau (red.) 2010.

Prinsloo, D.J., Taljard, E. and Goosen, M. *Optical Character Recognition and text cleaning in the indigenous*

South African languages. To appear: SpillPlusSenekal, B. A. and Kotzé, E. 2018. *Die ontwikkeling van 'n koste-effektiewe en byderwetse multimedia digitale argief by EPOG in Orania*. LitNET Akademies 15(3), 239 – 275.

Exploring Afrikaans word embeddings with analogies and nearest neighbours

Gaustad, Tanja
Eiselen, Roald
Centre for Text Technology (CTexT), North-West University, South Africa
Tanja.Gaustad@nwu.ac.za
Roald.Eiselen@nwu.ac.za

Abstract

This paper presents an exploration of word embeddings for Afrikaans using the analogies and nearest neighbours methodologies. We compare the results on three types of embeddings (fastText, FLAIR and GloVe) on a novel analogy data set for Afrikaans, inspired by the Bigger Analogy Test Set: BATS (Gladkova *et al.* 2016). Our analysis shows that for Afrikaans, similar to English, the types of embeddings influence the quality of analogies found for different linguistic tasks. Our investigation also demonstrates, however, that these Afrikaans embeddings do not encode as clear a linguistic representation as with English embeddings. The exact reason for this is subject to future work, but the added morphological complexity and the lack of data most likely play a role.

Keywords: Text embeddings, Afrikaans, Analogy, Evaluation, Low-resource languages

1 Introduction and background

Over the last decade there has been a fundamental shift in the field of natural language processing (NLP) with the broad adoption of deep neural networks (DNNs), leading to major advances across the field. Underpinning this shift has been the introduction of more sophisticated methods for representing language data in numerical form, specifically vectorised real value representations known as word embeddings. These representations are a prerequisite for applying deep learning techniques to various NLP technologies. At the same time these embeddings have removed a significant portion of the linguistics that formed part of the NLP development cycle (even with traditional machine learning techniques) and resulted in a now almost completely engineering and state-of-the-art driven pursuit.

One of the features of these more complex representations is that there is no clear human interpretable connection between the vectorised representations and existing linguistic knowledge. This in turn makes the machine learning components, which are already very complex and difficult to interpret, almost impossible to fully understand. Even so, developers have made broad claims about the linguistic information that is represented in these embeddings on both morphological, syntactic, and semantic levels (Mikolov *et al.* 2013a, Pennington *et al.* 2014). To support these claims, different tests have been designed with the aim of indirectly explaining the information that is contained in the vector representations, primarily for English. More recently, there have also been more linguistically motivated investigations to attempt to get a better understanding of the information encoded in these embeddings and whether there are correlations with existing linguistic concepts and knowledge (Allen & Hospedales 2019, Miaschi & Dell'Orletta 2020, Warstadt *et al.* 2019).

For Afrikaans, there have been a limited number of investigations into the use of deep learning and word embeddings (Hanslo 2021, Heyns & Barnard 2020, Loubser & Puttkammer 2020, Ralethe 2020, Van Heerden & Bas 2021), mostly focussing on the application of deep learning to various NLP tasks. Until recently there were only three freely available Afrikaans word embedding models (Conneau *et al.* 2020, Grave *et al.* 2018), all without any direct assessment of their quality. Most recently, Eiselen (2022) released five new embedding models for Afrikaans (freely available from [1]), trained on a larger curated data set, of which three will be used in this study.

To our knowledge there has not been an in-depth investigation into the nature of word embeddings for Afrikaans, and whether the tests and claims made for English embeddings hold for a language such as Afrikaans, which is morphologically more complex both in terms of derivation and inflection, but also very productive in terms of compounding, unlike

English. Afrikaans also has substantially less data available to train these embedding models.

With this background in mind, our study aims at an exploratory investigation of Afrikaans word embeddings for three different architectures (GloVe, fastText, and FLAIR), applying existing evaluation techniques to answer the following questions:

- Do different embedding models encode different types of information for more morphologically complex and less resourced languages, such as Afrikaans?
- Are the intrinsic evaluation methods for English applicable to more morphologically complex languages, such as Afrikaans?

The following section provides a short overview of word embeddings, the three architectures under consideration and the training data used. Section 3 gives an overview of word embedding analysis techniques and the experimental design for Afrikaans, followed by an analysis of the results for the different experiments in Section 4. We conclude the investigation in Section 5 with further discussion of our findings and areas for possible future work.

2 Embedding architectures and training procedures

Finding meaningful numerical representations for text, and especially words, has a long history in NLP (Pennington *et al.* 2014). This is especially true in the machine learning context where these representations are a requirement for the models to be trained. Although work on learning these types of representations has been ongoing since Bengio *et al.* (2003), the predominant approach to representing words in machine learning models was so-called one hot vectors, where each word in a vocabulary is represented by a sparse vector containing zeros for all positions except the one for the particular word, which is set to 1. This method was usable, but only included information about whether the word is a member of the vocabulary or not. This changed in 2013 with the introduction of word2vec (Mikolov *et al.* 2013a, Mikolov *et al.* 2013b, Mikolov *et al.* 2013c), where real-valued vectors are learned by a combination of sentence level cooccurrences and a log-linear classifier to generate an output vector of predefined length. This was followed shortly thereafter by another embedding technique, Global Vectors (GloVe) (Pennington *et al.* 2014). Both methods allowed for training on huge amounts of data efficiently and the learned vector representations resulted in improvements in many downstream NLP technologies when combined with various deep learning techniques.

One of the major shortcomings of these "classic" embedding models is that each word has a single embedding, irrespective of the context in which the word appears. This has been addressed by more recent embedding and language models that leverage different DNN architectures, such as convolutional, recurrent, and transformer neural networks. These models learn a model for generating a vector output, which can adapt the vector representation for a word by taking the context in which the word appears into account. This has further allowed for major gains in downstream NLP tasks, at the cost of at least one very important aspect, namely explainability.

From the outset of developing embeddings, it was clear that although the vector representations did correlate with several semantic and morpho-syntactic attributes of English, it was difficult to determine what the model is learning. The nature of the embeddings - large vectors of real-valued numbers - and their training procedures obfuscate the meaning of a particular value in a particular vector position and how the values correlate with linguistic attributes. This has become even worse with the use of DNNs to generate the representations, since there are so many variables in the process, that it becomes almost impossible to determine if there are specific linguistic attributes associated with specific vector positions or regions. Even though there have been several efforts to propose methods for investigating embeddings, there is still no clear methodology for investigating the quality of the embeddings and explaining the values associated with the models. Furthermore, most of these investigations have focussed on English exclusively, and little work has been done to determine how representations perform in linguistically different and/or less-resourced

environments. For this study we concentrate on three embedding architectures, two of the most common classical embeddings, and one recurrent neural network, namely fastText, GloVe, and FLAIR embeddings.

fastText (Bojanowski *et al.* 2017) is an extension of the original word2vec (Mikolov *et al.* 2013a) that includes character n-grams in the embedding calculations to ensure that previously unseen words also generate embeddings. GloVe embeddings (Pennington *et al.* 2014) differ slightly from fastText in that they use global cooccurrences of words to train a log-bilinear regression model for generating the embeddings, and only consider words. Both of these models generate a single embedding for a word, irrespective of the context of the word. FLAIR embeddings on the other hand train a long-short-term-memory recurrent neural network to generate a representation based on a character sequence. This has two advantages: i) the same word in different contexts can have different representations reflecting the context; and ii) because the model considers characters, and not words, as basic units, any sequence of characters will get an embedding, irrespective of whether it has been seen during training. This last characteristic is especially useful in less-resourced environments where data sparsity remains a major issue. Both fastText and FLAIR each have two flavours, but due to space constraints we will focus only on the fastText continuous bag-of-words (CBoW) and FLAIR backward models in our analysis.

The primary prerequisite for training any type of embedding is a large collection of text data, typically in the order of billions of words. Unfortunately, no such large data collection exists for Afrikaans. For the purposes of this study, we used a combination of freely available data, including NCHLT Afrikaans Text Corpora (Eiselen & Puttkammer 2014), Autshumato Afrikaans monolingual text data (Snyman *et al.* 2013), and Wikipedia [2], as well as in-house data sets with restricted access due to copyright. In total, the models were trained on approximately 250 million words, which is far less data than is typically used in learning embeddings for most of the well-resourced languages of the world.

Since the current study is primarily interested in exploring the characteristics of the vector representations, default settings were used for training each of the embeddings.

3 Analysis techniques for word embeddings: Experimental design for Afrikaans

As mentioned above, a purely intrinsic evaluation of word embeddings remains elusive as the vectors contain large numbers of numeric values that do not clearly correspond to specific linguistic features and are therefore not easily interpretable by humans. Word embeddings are usually evaluated when used as input to a larger system which then shows improved performance. With this type of extrinsic evaluation it is difficult, however, to assess the input from the embeddings to the overall performance compared to e.g. the architecture of the system (Schnabel *et al.* 2015). We will now discuss how we used existing analysis techniques to evaluate and explore Afrikaans embeddings.

3.1 Analogies

There have been various attempts to investigate how embeddings for different words correlate and to show that they represent some (type of) linguistic attribute (Allen & Hospedales 2019, Miaschi & Dell'Orletta 2020, Tulkens *et al.* 2016, Warstadt *et al.* 2019). One such technique is to use analogy-based data to test the identification of linguistic relations using word embeddings (Mikolov *et al.* 2013a, Turney 2012). The most cited analogy is undoubtedly "Which word is to king as woman is to man?" with the expected answer "queen".

Mikolov *et al.* (2013a) introduced the Google analogy test set for English which contains nine morpho-syntactic and five semantic categories. The semantic tasks are all encyclopaedic whereas the morpho-syntactic categories include two tasks on derivational morphology, six on inflectional morphology and one encyclopaedic task, with between 20 and 70 unique word pairs each. As has been noted by Gladkova *et al.* (2016), there are two issues with existing test sets: firstly, most of them are not balanced for different types of linguistic relations and secondly, results are usually reported as an

average over an entire test set and not per type of relation. To remedy the first shortcoming, they introduced the Bigger Analogy Test Set (BATS) covering four main types of linguistic relations: inflectional and derivational morphology as well as lexicographic and encyclopaedic semantics. Each main type in turn contains 10 different relations with 50 unique word pairs each.

To date, there have been limited investigations of embeddings for South African languages (Dlamini *et al.* 2021), and no such analogy test sets exist for Afrikaans specifically. For this initial exploration of Afrikaans word embeddings, BATS served as inspiration to create a small set of analogies. We did not include any lexicographic semantic tasks at this stage but decided to focus on inflectional and derivational morphology plus two encyclopaedic semantics tasks for comparison with English.

The first step was a careful analysis of the categories used: being based on English, not all of them are applicable to a different language. For instance, one of the inflectional morphology tasks in BATS, verb plural formation, is not present in Afrikaans. Furthermore, for categories that are applicable, simple translation is usually not a viable option due to differences in usage, frequencies, and formations of words in Afrikaans. For each category covered in our study, we attempted to get a representative sample of as many aspects of the category as possible. For plural nouns for instance, a substantial number of different classes of regular and irregular plurals found in textbooks and grammars were included. The same holds for comparative and superlative adjectives. One linguistic aspect that has had limited investigation in this kind of testing, but is very prevalent in Afrikaans, is compounding. Therefore, a very small set of noun compounds was included to determine how they are represented in the embeddings.

Our test set for Afrikaans includes two semantic tasks, both encyclopaedic, and 11 morpho-syntactic tasks, three derivational, seven inflectional as well as compounding. Overall, there are 16,313 analogy "questions". Table 1 shows an overview of the categories chosen, including how many word pairs per task and an example for each.

Answers to analogy questions are calculated by taking the vector representation of word 1 (V_{w1}), subtracting the vector of word 2 (V_{w2}), related either semantically or morpho-syntactically, then adding the vector of a third word (V_{w3}). The resulting vector (V_{result}) is then compared to the vectors of all words in the model to find the vector(s) with the smallest Euclidean distance. The expectation is that the nearest vector to the result vector will express the same relationship to W3 as the relationship between W1 and W2. The prototypical example, $V_{king} - V_{man} + V_{woman}$ should result in a vector that has the smallest distance to V_{queen}. Similarly, $V_{stronger} - V_{strong} + V_{clear}$ should

Table 1: Analogy data set for Afrikaans: Types of linguistic relations, number of unique word pairs and examples.

Category	Subcategory	Task	# word pairs	Example
Morpho-syntactic	Derivational	Noun to Verb (*be-*, *ver-*)	20	man – beman taal – vertaal
		Noun to Adj (*-ies*)	10	simbool - simbolies
		Verb to Adj (*-baar*)	10	lees – leesbaar
	Inflectional	Adj comparative	41	duur – duurder
		Adj superlative	41	duur – duurste
		Adj comparative to superlative	41	duurder – duurste
		Attributive *-e*	10	teoreties – teoretiese
		Noun diminutive	56	hand – handjie
		Noun plural (reg/irreg)	76	kop – koppe
		Verb past tense	32	doen – gedoen
	Compounding	Noun compounding	19	landbousektor
Semantic	Encyclopaedic	Country - Capital	23	Duitsland – Berlyn
		Man - Woman	28	buurman – buurvrou

result in a vector closest to $V_{clearer}$. The vector visualisation in Figure 1 provides an intuition for why this should work. The offsets between man and woman, and king and queen, although not exactly the same, are similar. Therefore, removing the man characteristics from king, and adding woman's characteristics, should yield a vector in close proximity to queen.

3.2 Nearest neighbours

A second method described by Collobert & Weston (2008) and also referenced by Mikolov *et al.* (2013a) is the analysis of the nearest neighbours for a specific set of words. The nearest neighbour of a word is again determined by finding those word vectors which have the smallest Euclidean distance between the vector for the query and vectors for any other words for which embeddings exist in the model. The hypothesis is that a qualitative review of the neighbours provides additional insight into the types of relationships that the embeddings are learning. As an example, the fastText English embedding for the word "run" includes "runs, running, ran" which indicates the encoding of some morpho-syntactic properties. Although it is not possible to create a single metric for evaluation purposes, it is a useful procedure to gain an understanding of the underlying information that is encoded in the embeddings, such as hypernymy, hyponymy, synonymy, or morpho-syntactic relations.

One of the caveats to keep in mind with the nearest neighbour analysis is that different types of relations may be found within a single

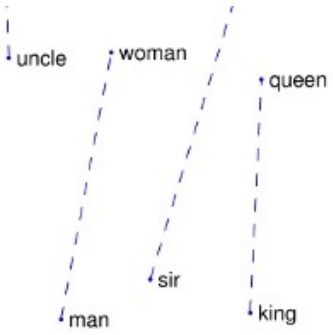

Figure 1: Gender related vector representation (from Pennington *et al.* (2014))

embedding architecture, and it may not always be immediately obvious what information is encoded in the embeddings. Consistent patterns can be found but should only be used to draw very broad and general conclusions about the encoded information.

To investigate the information encoded in the different embedding architectures for Afrikaans, we selected two words from each category listed in Table 1, 26 in total, and generated the five nearest neighbours for each word in each of the different architectures. These were then manually reviewed to determine the quality and nature of the embeddings.

3.3 Downstream task evaluation

The most common method for validating the quality of word embeddings is their application as part of a downstream task, such as POS tagging, named entity recognition or question answering. The use of embeddings rather than one hot encoding was one of the first steps enabling the current deep learning trend in NLP, and it has been consistently shown that using embeddings in downstream tasks improves the quality of the technology. This has also been shown to be the case for Afrikaans where a combination of FLAIR embeddings improve both POS tagging and NER results over previous models (Eiselen 2022). Due to space constraints we do not include these results in the current analysis.

4 Analysing Afrikaans word embeddings: Results

One of our initial motivations for this research was to investigate whether the type of information encoded in different embedding models is similar for linguistically different languages and whether methods used to evaluate English embeddings are also applicable to morphologically more complex languages. We will first present the results for the analogy task per linguistic (sub)task, including a thorough discussion of our observations. This intrinsic quantitative evaluation of the word embeddings for Afrikaans is followed by an intrinsic qualitative analysis investigating the nearest neighbours as described in section 3.2.

4.1 Analogies: Quantitative evaluation

Using the analogy data set for Afrikaans described earlier, the accuracy for each type of task is calculated separately. Previous studies mostly report only the accuracy of the word with the smallest Euclidean distance. Our evaluation, however, includes two accuracy scores: one for matches from the closest word (position 1) and one for matches from words in the subsequent four positions (position 2-5). Including more than only the closest matches gives us more insight into the different embedding representations for the various analogy tasks. Furthermore, when calculating accuracies, all input question words were excluded from the results. Omitting this adaptation resulted in much worse results (an effect also noted in Linzen (2016)). Table 2 shows the results for the different linguistic categories (aggregated at subcategory level) and types of embeddings.

In our experiments for Afrikaans, the overall best performing task and embedding type is GloVe on the semantic tasks with 51,11% accuracy, whereas the worst results are also obtained with GloVe, but on the derivational morphology tasks (1,11%). Compared to results for English on the Google data set (ranging from nearly 60% (Mikolov *et al.* 2013a) to high 60% (Levy & Goldberg 2014)), it is noteworthy how poorly all the embedding types perform on all of the tasks for Afrikaans. Contrasting our GloVe outcomes with Gladkova *et al.*'s (2016) more detailed results on BATS, the performance on Afrikaans is again quite a bit lower.

Focussing on the type of tasks, for the derivational tasks FLAIR performs best and GloVe worst. Both GloVe and fastText embeddings have a very high percentage of words not found in the top ten (87% and 68% respectively) which explains their poorer performance. The likely reason for the poor performance on derivations is the fact that per definition the paired words belong to different syntactic categories and typically do not appear in similar positions, hence do not have similar co-occurrences to the query word and will therefore have substantially different vector representations.

For inflectional morphology, fastText has the highest percentage of correct analogies for the first position, while GloVe has the lowest, although the difference in accuracy is fairly small compared to the other tasks. Interestingly enough, there are marked differences in correct words found in positions 2-5: FLAIR finds the searched for analogy in more than 40% of the cases, whereas the other embedding types only find it in slightly more than 20%. The FLAIR embeddings also find most analogies whereas GloVe finds the least. This can be explained by the fact that FLAIR embeddings encode character sequences and typically inflectional morphology happens at the character level. With regard to the subtasks for inflection, plural and diminutive forms are hardest to detect.

The results for the compound analogies indicate that the word embeddings do not learn a

Table 2: Accuracy for the Afrikaans analogy test set on different linguistic tasks (aggregated at subcategory level) for three word embedding types (best performance in position 1 per task type in bold).

Task Type	Architecture	Position 1	Positions 2-5	Not found
Derivational	fastText	11,94%	14,44%	68,06%
Derivational	FLAIR	**25,28%**	33,33%	33,06%
Derivational	GloVe	1,11%	9,44%	86,94%
Inflectional	fastText	**26,56%**	22,82%	41,51%
Inflectional	FLAIR	22,89%	41,45%	25,69%
Inflectional	GloVe	22,32%	20,98%	52,37%
Compounds	fastText	0,00%	0,00%	94,74%
Compounds	FLAIR	0,00%	0,00%	89,47%
Compounds	GloVe	0,00%	0,00%	100,00%
Semantic	fastText	16,56%	25,83%	48,10%
Semantic	FLAIR	7,92%	14,42%	72,58%
Semantic	GloVe	**51,11%**	33,20%	12,76%

representation of the constituents of the compound. Performing an analogy test that isolates the head of the compound results in a representation that is in a completely unrelated vector space, with no correlation to either the compound or its head. Although compounds are less frequent than the compound head in general, this does not seem to be the main contributing factor to the poor performance. As is discussed in the following section, the nearest neighbours of the head do contain many compounds, even relatively low frequency compounds, indicating that the full compound is seen as similar to the head, but not necessarily on a constituent level. This aspect of embeddings has not been studied extensively and will require further investigation in future.

The results for the semantic analogy tasks are the reverse of the morpho-syntactic ones (excluding compounding): GloVe very clearly outperforms all the other embedding types on all measurements. Here, the FLAIR embeddings perform the worst, also on all accounts. The one caveat to these results is that FLAIR embeddings are by nature contextual, and different vector representations will be generated when considering the words in sentence contexts, which was not the case in our tests. It may well be that the FLAIR embeddings perform better on semantic analogy tasks if vectors are generated for words in a sentence context. Unfortunately, there is not currently a well-defined methodology for generating embeddings for this type of task and it is something that will need to be considered in future work, especially if this type of analysis is undertaken with other types of representations, such as transformer models.

To summarize, our results corroborate earlier findings on English that different types of embeddings work best for different linguistic analogy tasks. In addition, our results on this analogy test set indicate that inflectional morphology is easier to model than derivational morphology, whereas compounding, a typical feature of Afrikaans, is very difficult to model. Overall, performance on Afrikaans, a more morphologically complex and productive language, is poorer than expected.

4.2 Nearest neighbours: Qualitative evaluation

After the more quantitative analysis using analogies, we now examine the nearest neighbours for Afrikaans, whether they differ from our expectations, and what we can learn from this examination.

fastText embeddings

The main finding for the fastText embeddings is that there are little to no examples of semantic relations in neighbours for any of the words selected, and in almost all cases the query is a substring within the set of nearest neighbours, see e.g. for *verdeel* (divide) and *Berlyn* (Berlin):

> *verdeel* ▭ *opverdeel, onderverdeel, onverdeel, verdeelpyp, verdeelbaar*
>
> *Berlyn* ▭ *Berlyn-Schönefeld, Berlyner, Berlynse, Berlynmuur, Wes-Berlyn.*

This can primarily be attributed to the fact that the inclusion of subword information in the embeddings has a strong effect on the vector representations and coincides with the fact that the morphology of Afrikaans is more productive than English, both in terms of inflectional and derivational paradigms. The consequence of this is that any inflectional or derivational form exhibiting some form of typographic change, e.g. shortening of the double vowels in plural forms, are not typically associated with the query word and therefore not returned as nearest neighbour. Furthermore, Afrikaans being a compounding language means that a large number of words closely associated with a query tend to be either inflections of the query or a compound including the query, rather than semantically related words as is often the case in English.

FLAIR embeddings

As was previously shown in Section 4.1, and expected given the evaluation parameters, there are essentially no semantic relationships between the query words and nearest neighbours for the FLAIR embeddings. Unlike the fastText embeddings, the FLAIR embeddings do not include the query term as a substring of the neighbours, but there are strong correlations with inflectional patterns. As an example, the nearest

neighbours for the word *leesbaar* (readable) are as follows:

> *leesbaar* ▭ *leeservaring, leefwêreld, kwesbaar, leefwyse, leesstof, leefstyl, vloeibaar, voorspelbaar, aanpasbaar*

From this set we see that the model either agrees with the 'lee' substring at the beginning of the word or the *-baar* (-able) morpheme at the end. This is an indication that the model is more likely to model affix structure.

GloVe embeddings

The embeddings for GloVe are substantially different from the other types, with a combination of morpho-syntactic, semantic, and cooccurrence instances showing up in the list of nearest neighbours, for example:

> *hoog* ▭ *hoë, laag, bo, bokant, hoogte, hoër, ver, meter, so*

> *ironie* ▭ *humor, satire, sarkasme, simboliek, tikkie, ironiese, tragiese*

In the examples for *hoog* (high), there are inflections - *hoë* (high), *hoogte* (height), *hoër* (higher); semantically related words - *laag* (low), *bo* (above), *bokant* (above, top); as well as words that frequently cooccur with *hoog* - *meter hoog* (meter's high), *so hoog* (so high). These cooccurences are not necessarily the most frequent as *te* (too), *is* (is) and *baie* (very) all occur more frequently with *hoog* than *meter* (VivA 2022). The same pattern also occurs for *ironie* (irony) with all three types of relations found in the nearest neighbours.

5 Discussion and future work

Our explorations of Afrikaans embeddings have shown that, similar to other languages, different types of embeddings work best for different linguistic analogy tasks. However, a careful analysis of the analogies and nearest neighbours results also demonstrates that these embeddings do not encode as clear a linguistic representation as for English. There are two possible reasons for these differences: Afrikaans is linguistically different to a relevant degree or more data is needed to train more representative embeddings. Currently, we do not know what the main source of the shortcomings for Afrikaans embeddings is, but surmise that most likely both the added morphological complexity and the lack of data have an influence.

In the case of linguistic diversity/complexity, this would mean the more different a language is compared to English, e.g. other South African languages such as isiZulu or Setswana, the more carefully word embeddings should be trained and the more critically they have to be evaluated. If data sparsity is the culprit (even though 250 million is middle-ground in terms of resources), this does not bode well for resource-scarce languages when building and subsequently using embeddings for NLP tasks and/or trying to understand what they represent.

Overall, as embeddings for morphologically complex, compounding languages are substantially different to English, both how to train these embeddings as well as how we analyse them need to be rethought, especially for under-resourced languages.

Future work includes building a full analogy set covering more linguistic categories relevant for Afrikaans. Expanding these explorations to the other South African languages is also an interesting challenge, especially given their high morphological productivity in conjunction with very little data.

Notes

[1] https://repo.sadilar.org

[2] https://dumps.wikimedia.org/

Acknowledgements

This work was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Sports, Arts and Culture.

References

Allen, C & Hospedales, T 2019, 'Analogies explained: Towards understanding word embeddings', *In*: Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, PMLR, pp. 223-231.

Bengio, Y, Ducharme, R, Vincent, P & Jauvin, C 2003, 'A neural probabilistic language model',

Journal of Machine Learning Research, vol. 3**,** pp. 1137–1155.

Bojanowski, P, Grave, E, Joulin, A & Mikolov, T 2017, 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics,* vol. 5**,** pp. 135-146.

Collobert, R & Weston, J 2008, 'A unified architecture for natural language processing: Deep neural networks with multitask learning', *In*: Proceedings of the 25th International Conference on Machine learning, Valencia, Spain, pp. 160-167.

Conneau, A, Khandelwal, K, Goyal, N, Chaudhary, V, Wenzek, G, Guzmán, F, Grave, É, Ott, M, Zettlemoyer, L & Stoyanov, V 2020, 'Unsupervised Cross-lingual Representation Learning at Scale', *In:* Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, ACL, pp. 8440-8451.

Dlamini, S, Jembere, E, Pillay, A & van Niekerk, B 2021, 'isiZulu Word Embeddings', *In:* Proceedings of the 2021 Conference on Information Communications Technology and Society, Durban, South Africa, IEEE, pp. 121-126.

Eiselen, R 2022, 'Afrikaans Text Embeddings for Sequence Labelling with Deep Neural Networks', *In:* Proceedings of the Southern African Conference for Artificial Intelligence Research 2022, Stellenbosch, South Africa, SACAIR.

Eiselen, R & Puttkammer, MJ 2014, 'Developing Text Resources for Ten South African Languages', *In:* Proceedings of the 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, ELRA, pp. 3698–3703.

Gladkova, A, Drozd, A & Matsuoka, S 2016, 'Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't', *In:* Proceedings of the NAACL Student Research Workshop, San Diego, CA, ACL, pp. 8-15.

Grave, É, Bojanowski, P, Gupta, P, Joulin, A & Mikolov, T 2018, 'Learning Word Vectors for 157 Languages', *In:* Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan, ELRA, pp 3483-3487.

Hanslo, R 2021, 'Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages', *In:* Proceedings of the 16th Conference on Computer Science and Intelligence Systems (FedCSIS), Virtual, IEEE, pp. 115-119.

Heyns, N & Barnard, E 2020, 'Optimising word embeddings for recognised multilingual speech', *In:* Proceedings of the Southern African Conference for Artificial Intelligence Research Conference 2020, Online, Virtual, SACAIR, pp. 102-116.

Levy, O & Goldberg, Y 2014, 'Linguistic Regularities in Sparse and Explicit Word Representations', *In:* Proceedings of the 18th Conference on Computational Language Learning, Baltimore, MD, ACL, pp. 171-180.

Linzen, T 2016, 'Issues in evaluating semantic spaces using word analogies', *In:* Proceedings of the 1st Workshop on Evaluating Vector Space Representations for NLP, Berlin, Germany, ACL, pp. 13-18.

Loubser, M & Puttkammer, MJ 2020, 'Viability of Neural Networks for Core Technologies for Resource-Scarce Languages', *Information,* vol. 11**,** pp. 41-57.

Miaschi, A & Dell'Orletta, F 2020, 'Contextual and non-contextual word embeddings: an in-depth linguistic investigation', *In:* Proceedings of the 5th Workshop on Representation Learning for NLP, Seattle, WA, ACL, pp. 110-119.

Mikolov, T, Chen, K, Corrado, G & Dean, J 2013a, 'Efficient estimation of word representations in vector space', *In:* Proceedings of the International Conference on Learning Representations 2013, Scottsdale, AZ.

Mikolov, T, Sutskever, I, Chen, K, Corrado, GS & Dean, J 2013b, 'Distributed representations of words and phrases and their compositionality', *Advances in Neural Information Processing Systems,* vol. 26**,** pp. 3111–3119.

Mikolov, T, Yih, W & Zweig, G 2013c, 'Linguistic regularities in continuous space word

representations', *In:* Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, GA, ACL, pp. 746–751.

Pennington, J, Socher, R & Manning, CD 2014, 'Glove: Global vectors for word representation', *In:* Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, ACL, pp. 1532–1543.

Puttkammer, MJ, Eiselen, R, Hocking, J & Koen, F 2018, 'NLP Web Services for Resource-Scarce Languages', *In:* Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Melbourne, Australia, ACL, pp. 43–49.

Ralethe, S 2020, 'Adaptation of deep bidirectional transformers for Afrikaans language', *In:* Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, ELRA, pp. 2475-2478.

Schnabel, T, Labutov, I, Mimno, D & Joachims, T 2015, 'Evaluation methods for unsupervised word embeddings', *In:* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, ACL, pp. 298-307.

Snyman, DP, McKellar, CA & Groenewald, H 2013, 'Autshumato English-Afrikaans Parallel Corpora', Data set, v1.0, South African Centre for Digital Language Resources (SADiLaR), https://hdl.handle.net/20.500.12185/397.

Tulkens, S, Emmery, C & Daelemans, W 2016, 'Evaluating Unsupervised Dutch Word Embeddings as a Linguistic Resource', *In:* Proceedings of the 10th International Conference on Language Resources and Evaluation, Portorož, Slovenia, ELRA, pp. 4130-4136

Turney, PD 2012, 'Domain and function: A dual-space model of semantic relations and compositions', *Journal of Artificial Intelligence Research,* vol. 44, pp. 533-585.

Van Heerden, I & Bas, A 2021, 'AfriKI: Machine-in-the-Loop Afrikaans Poetry Generation', *In:* Proceedings of the 1st Workshop on Bridging Human-Computer Interaction and Natural Language Processing, Virtual, ACL, pp. 74-80.

VivA 2022, 'Korpusportaal: Omvattend 1.11', *Virtuele Instituut vir Afrikaans (VivA)*, viewed 1 September 2022, <https://viva-afrikaans.org>.

Warstadt, A, Cao, Y, Grosu, I, Peng, W, Blix, H, Nie, Y, Alsop, A, Bordia, S, Liu, H & Parrish, A 2019, 'Investigating BERT's Knowledge of Language: Five Analysis Methods with NPIs', *In:* Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Hong Kong, China, ACL, pp. 2877-2887.

Deriving lexical statistics for psycholinguistic research on isiXhosa

Berghoff, Robyn
Stellenbosch University
berghoff@sun.ac.za

Abstract

Psycholinguistic research on isiXhosa and related Bantu languages is scarce. For research on lexical processing in particular, a prerequisite is data on lexical properties that impact word recognition, such as word frequency and neighbourhood density. This paper describes the derivation of these and related lexical statistics from a newly created 4.8-million-word isiXhosa corpus. It then reviews the potential applications of such a lexical database for research on language acquisition, language development, and language processing. The paper closes with recommendations for further work in this domain.

Keywords: isiXhosa, psycholinguistics, lexical statistics, lexical processing, corpus

Introduction

The vast majority of language acquisition and processing research focuses on a small subset of the world's languages, specifically those from the Germanic and Romance branches of the Indo-European language family (Bylund, Khafif & Berghoff 2022; Norcliffe, Harris & Jaeger 2015). Within the neglected language groups, the Bantu languages are particularly understudied. This narrow focus in terms of language typology severely limits the generalizability of theories of language processing.

isiXhosa is an agglutinating language, meaning that its words typically consist of multiple morphemes that are concatenated in a relatively transparent manner. It has two particular features that distinguish it from more commonly studied agglutinating languages such as Turkish, Basque, and Hungarian (see van de Velde et al. 2019 for discussion). Firstly, alongside suffixation, it makes widespread use of prefixation to produce morphologically complex words. Secondly, it has a rich grammatical gender or noun class system, whereby nouns are divided into 15 groups, with noun class agreement being marked on several syntactic constituents (e.g., verbs, adjectives/relatives, determiners). As language-specific properties of affixation (e.g., Boudelaa & Marslen-Wilson 2011) and grammatical gender (e.g., Colé, Pynte & Andriamamonjy 2003) are known to affect language processing, psycholinguistic examinations of isiXhosa and related languages have much to contribute to theories of lexical and morphological processing. Further, in terms of practical applications, an understanding of how language and literacy development proceeds in languages of this sort is indispensable in designing teaching and intervention materials for young learners (Pretorius 2019).

A sine qua non of robust psycholinguistic research into lexical processing are statistics on certain lexical properties known to influence word recognition, such as word frequency and neighbourhood density. Databases of such statistics are increasingly being developed and made available to facilitate research on more commonly studied languages. Examples include GreekLex, for Greek (Ktori, van Heuven & Pitchford 2008); EsPal, for Spanish (Duchon et al. 2013); Aralex, for Modern Standard Arabic (Boudelaa & Marslen-Wilson 2010); StimulStat, for Russian (Alexeeva, Slioussar & Chernova 2018); the Chinese Lexical Database, for Mandarin (Sun et al. 2018); P-PAL, for Portuguese (Soares et al. 2018); and E-Hitz, for Basque (Perea et al. 2006); as well as CLEARPOND (Cross-Linguistic Easy Access Resource for Phonological and Orthographic Neighbourhood Densities; Marian et al. 2012), which provides lexical data for five widely examined European languages.

Resources on African indigenous languages are scarce. Unsurprisingly, then, the kinds of data needed for robust psycholinguistic research on

isiXhosa processing and acquisition are lacking. This paper describes the generation of lexical statistics for isiXhosa. It identifies and defines the types of lexical statistics needed for lexical processing research on isiXhosa and exemplifies their calculation based on a newly created 4.8-million-word isiXhosa corpus. The paper concludes by reviewing potential applications of such a database and outlining steps for future work.

Characteristics to be included in a database of lexical statistics for isiXhosa

This section reviews the characteristics that should, at a minimum, be included in a database of lexical statistics for isiXhosa. This overview focuses on characteristics that are relevant specifically to visual word recognition.

Frequency

Frequency is arguably the most important variable in studies of lexical processing (van Heuven et al. 2014), where more frequent words are processed more rapidly than less frequent words. This effect can be explained on the basis of lexical activation, whereby more frequently encountered words have higher resting activation levels and are thus accessed more quickly than their low-frequency counterparts. A word's frequency is calculated based on the number of its occurrences in a corpus. It can be expressed on the standardized Zipf frequency scale, where the lower half of the scale (1–3) represents low-frequency words and the upper half of the scale (4–6) represents high-frequency words (words with a Zipf frequency above 7 tend to be function words). Zipf frequency is calculated as log10 (frequency per million words) + 3 (van Heuven et al. 2014).

Word length

Word length is calculated simply as the number of letters in a given word. At least in English, it has been found to have non-linear effects on word recognition independently of other variables such as number of syllables (New et al. 2006).

Neighbourhood statistics

A neighbour of a given word is any word that can be created by substituting, adding, or deleting a single letter (for example, the isiXhosa *ubisi* 'milk' has as neighbours *ubusi* 'honey' and *usisi* 'sister', among others). The neighbourhood density of a word is equal to the number of its neighbours. Neighbourhood density effects on lexical processing typically manifest as processing slowdowns for words with more neighbours (Andrews 1997), which is attributed to the fact that when recognizing a word with many neighbours, numerous candidate lexical items become activated and must consequently be inhibited for the correct item to be selected. A related variable that is also of importance is neighbourhood frequency, which is the average frequency of a given word's neighbours. Here, a word with high-frequency neighbours takes longer to recognize than a word with low-frequency neighbours (e.g., Brysbaert, Mandera & Keuleers 2018).

Method

The corpus

The calculation of lexical statistics requires a sizeable corpus of contemporary language materials. The corpus used in this paper was created by combining the isiXhosa corpora provided in the Leipzig Corpora Collection (Goldhahn, Eckart & Quasthoff 2012) with a new corpus created by the author from the online isiXhosa newssite *Isolezwe lesiXhosa*. The Leipzig isiXhosa corpora consist of texts randomly collected from the web and Wikipedia and therefore cover a multitude of topics (see Goldhahn, Eckart & Quasthoff 2012 for discussion). The *Isolezwe* corpus, on the other hand, contains reports on general news, sports, entertainment, opinion, and agriculture. This corpus was created via web-scraping using the rvest package (version 1.0.2; Wickham 2021) in the R environment for statistical computing (version 4.2.1; R Core Team 2022). The entire history of articles that was available from the

site's inception (26 June 2015) up until 24 June 2022 was scraped.

To create the final corpus, each subcorpus was read into R and subjected to basic cleaning (e.g., removal of digits) using the stringr package (version 1.4; Wickham 2019). Tokenization of each subcorpus was then performed using the "unnest_tokens" function from the tidytext package (version 0.3.3; Silge & Robinson 2016).

Details of each component of the final corpus are provided in Table 1.

Table 1: Components of final corpus

Name	Tokens
Leipzig 2013 corpus	400,323
Leipzig 2015 corpus	153,661
Leipzig 2016 corpus	424,146
Leipzig 2017 corpus	343,517
Leipzig 2018 corpus	443,931
Leipzig 2020 corpus	436,772
Isolezwe corpus	2,656,625
Total	4,858,975

All the subcorpora were combined prior to further processing. The final corpus contained 466,957 distinct tokens. This size is comparable to that used in the calculation of lexical statistics for other languages (e.g., Basque; Perea et al. 2006).

Calculation of statistics

Frequency numbers were obtained using the tidytext package in R. These raw numbers were then converted to Zipf frequencies. The other lexical statistics were calculated using the LexiCAL program (Chee et al. 2021). This Windows application allows the user to input a corpus file, which specifies the tokens and their frequency in the corpus, for any alphabetic language. It then calculates the selected metrics and outputs the results to a separate file. For neighbourhood statistics, it also provides the neighbours of the words that are included in the corpus.

Excerpts from the database

This section presents excerpts from the database. To begin with, Table 2 lists the 20 most frequent words in the corpus with their raw and Zipf frequencies.

Table 2: Twenty most frequent words in the corpus

Word	Raw freq.	Zipf freq.
ukuba	73,396	7.18
le	22,958	6.67
xa	22,712	6.67
emva	21,155	6.64
kwaye	19,588	6.61
ke	19,169	6.60
okanye	18,274	6.58
lo	16,624	6.53
kodwa	16,387	6.53
kuba	15,903	6.51
uthi	14,762	6.48
nto	14,519	6.48
abantu	13,533	6.44
kunye	12,742	6.42
kakhulu	11,809	6.39
kule	11,120	6.36
afrika	11,014	6.36
ukuze	10,996	6.35
utshilo	10,806	6.35
into	9,845	6.31

Unsurprisingly, the majority of the 20 most frequent words are function words, with the exception of *uthi* 'you/he/she says', *(i)nto* 'thing',

abantu 'people', *kakhulu* 'very, a lot', *afrika* 'Africa', and *utshilo* 'you/he/she said'.

The crucial factor in designing psycholinguistic experiments, however, is not the raw frequency of items, but matching frequency and other lexical properties across items. Table 3 presents example database entries for ten randomly selected items from the corpus. In creating an experiment, the aim would be to select lexical items that are as closely matched on the numerical values in Table 3 as possible.

Applications

A database of lexical statistics such as that described in this paper has numerous applications for research on lexical processing, which requires careful control of word-level properties such as frequency and neighbourhood density. There is, for example, significant interest in whether recognition of morphologically complex words takes place at the whole-word level or whether it entails breaking a word down into its constituent morphemes. To the best of the author's knowledge, only one study has investigated this question in relation to a Bantu language (Setswana; Ciaccio, Kgolo & Clahsen 2020). The extent to which morphosyntactic processing differs across first- and second-language speakers of a language is also theoretically important and critically understudied in relation to Bantu languages (Spinner 2011).

Another set of applications arises in the domain of language development. Lexical databases of the type described in this paper have been developed specifically for use in psycholinguistic studies of children's language processing and for evaluating literacy materials aimed at developing readers (e.g., Corral, Ferrero & Goikoetxea 2009; Masterson et al. 2010; Schroeder et al. 2015; Terzopoulos et al. 2017). For such applications, corpora are typically compiled using a large selection of materials created specifically for children in order to increase the likelihood of children having been exposed to the language it contains.

Table 3: Example database entries

	Word length	Raw freq.	Zipf freq.	Neighbourhood (N) size	Example neighbours	N. freq. (mean)	N. freq (SD)
umntu	5	7,530	6.19	16	mntu, kumntu	212.37	569.68
amanzi	6	1,672	5.54	18	yamanzi, abanzi	70.05	88.16
ububele	7	46	3.98	11	ubuyele, ubukele	27.09	42.35
isakhiwo	8	135	4.44	7	izakhiwo, esakhiwo	52.86	89.6
ukulala	7	108	4.35	23	ukudlala, ukuhlala	131.35	268.59
ukubhala	8	508	5.02	18	ukubala, ukubhalwa	30.33	50.63
ukucinga	8	199	4.61	18	ukujinga, akucinga	8.15	7.21
ukuthengis	11	152	4.49	5	ukuthengiswa, ukumthengisa	36.2	50.77

a							
kaninzi	7	48	3.99	9	baninzi, maninzi	111.78	173.47
phakathi	8	4,730	5.99	10	ephakathi, iphakathi	110.5	130.46

Limitations and suggestions for further work

There are several additional steps that can be taken to improve on and expand the database presented in this paper. For one, after compiling the corpus and before processing it to derive lexical statistics, it would be desirable to cross-reference the corpus word list with a word list from an official isiXhosa dictionary. This cross-referencing process would enable misspellings to be filtered out from the corpus, thus reducing the number of spurious neighbours identified, and also facilitate the removal of non-isiXhosa words. At the time of writing, no such dictionary word list could be obtained in a digital format, and so this step has not yet been taken.

Another notable consideration is that all of the above calculations were based on word forms rather than roots or lemmas. This means that instead of, for example, *-lala* being treated as a lemma that surfaces in *ukulala*, *uyalala*, *siyalala*, and so forth, each of these word forms is treated as an individual item. This can also lead to inflation of neighbourhood density (however, the words affected by this issue – most notably, verbs – will tend to have their neighbourhood density inflated to the same extent). It remains an empirical question whether it is properties of the word form or the lemma that are better predictors of, for example, word recognition latency in languages such as isiXhosa. In order to address this question, lemmas could be obtained from corpus data using a morphological analyzer (e.g., du Toit & Puttkammer 2021).

Lastly, the work presented here could also be expanded by deriving phonological statistics for isiXhosa, such as syllable number and phonological neighbourhood density. Such statistics can be obtained via LexiCAL if each word entry is paired with a phonetic transcription and would enable research on spoken word processing in the language.

Conclusion

Psycholinguistic techniques that can capture language processing as it unfolds in real time have yet to be leveraged to examine the processing of isiXhosa and related languages. This paper has discussed one kind of resource – a database of lexical statistics on the language, compiled based on a large-scale corpus – that is necessary to address this research gap and realize the considerable theoretical and practical benefits of doing so. Future collaboration between (computational) linguists and language specialists will allow for the refinement of this resource and the creation of others.

Acknowledgements

This work was made by possible by an NRF Thuthuka grant (grant no. 138180) awarded to R. Berghoff.

References

Alexeeva, S, Slioussar, N & Chernova, D 2018, 'StimulStat: A lexical database for Russian', *Behavior Research Methods*, vol. 50, no. 6, pp. 2305–15.

Andrews, S 1997, 'The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts', *Psychonomic Bulletin & Review*, vol. 4, no. 4, pp. 439–61.

Boudelaa, S & Marslen-Wilson, WD 2010, 'Aralex: A lexical database for Modern Standard Arabic', *Behavior Research Methods*, vol. 42, no. 2, pp. 481–7.

Boudelaa, S & Marslen-Wilson, WD 2011, 'Productivity and priming: Morphemic decompo-

sition in Arabic', *Language and Cognitive Processes*, vol. 26, 4-6, pp. 624–52.

Brysbaert, M, Mandera, P & Keuleers, E 2018, 'The word frequency effect in word processing: An updated review', *Current Directions in Psychological Science*, vol. 27, no. 1, pp. 45–50.

Bylund, E, Khafif, Z & Berghoff, R 2022, 'Linguistic and geographic diversity (or lack thereof) in research on second language acquisition and multilingualism', *Applied Linguistics*. Manuscript submitted for publication.

Chee, QW, Chow, KJ, Goh, WD, Yap, MJ & Miwa, K 2021, 'LexiCAL: A calculator for lexical variables', *PloS one*, vol. 16, no. 4, pp. e0250891.

Ciaccio, LA, Kgolo, N & Clahsen, H 2020, 'Morphological decomposition in Bantu: a masked priming study on Setswana prefixation', *Language, Cognition and Neuroscience*, vol. 35, no. 10, pp. 1257–71.

Colé, P, Pynte, J & Andriamamonjy, P 2003, 'Effect of grammatical gender on visual word recognition: Evidence from lexical decision and eye movement experiments', *Perception & Psychophysics*, vol. 65, no. 3, pp. 407–19.

Corral, S, Ferrero, M & Goikoetxea, E 2009, 'LEXIN: A lexical database from Spanish kindergarten and first-grade readers', *Behavior research methods*, vol. 41, no. 4, pp. 1009–17.

du Toit, Jakobus S. & Puttkammer, MJ 2021, 'Developing core technologies for resource-scarce Nguni languages', *Information*, vol. 12, no. 12, p. 520.

Duchon, A, Perea, M, Sebastián-Gallés, N, Martí, A & Carreiras, M 2013, 'EsPal: One-stop shopping for Spanish word properties', *Behavior Research Methods*, vol. 45, no. 4, pp. 1246–58.

Goldhahn, D, Eckart, T & Quasthoff, U 2012, 'Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages', *Proceedings of the 8th International Language Resources and Evaluation*.

Ktori, M, van Heuven, W & Pitchford, NJ 2008, 'GreekLex: A lexical database of Modern Greek', *Behavior Research Methods*, vol. 40, no. 3, pp. 773–83.

Marian, V, Bartolotti, J, Chabal, S, Shook, A & White, SA 2012, 'CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities', *PloS one*, vol. 7, no. 8, pp. e43230.

Masterson, J, Stuart, M, Dixon, M & Lovejoy, S 2010, 'Children's printed word database: Continuities and changes over time in children's early reading vocabulary', *British Journal of Psychology*, vol. 101, no. 2, pp. 221–42.

New, B, Ferrand, L, Pallier, C & Brysbaert, M 2006, 'Reexamining the word length effect in visual word recognition: New evidence from the English Lexicon Project', *Psychonomic Bulletin & Review*, vol. 13, no. 1, pp. 45–52.

Norcliffe, E, Harris, AC & Jaeger, TF 2015, 'Cross-linguistic psycholinguistics and its critical role in theory development: early beginnings and recent advances', *Language, Cognition and Neuroscience*, vol. 30, no. 9, pp. 1009–32.

Perea, M, Urkia, M, Davis, CJ, Agirre, A, Laseka, E & Carreiras, M 2006, 'E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque)', *Behavior Research Methods*, vol. 38, no. 4, pp. 610–5.

Pretorius, E 2019, 'Getting it right from the start', in N Spaull & J Comings (eds), *Improving Early Literacy Outcomes*, BRILL, pp. 63–80.

R Core Team 2022, *R: A language and environment for statistical computing*, <https://www.R-project.org/>.

Schroeder, S, Würzner, K, Heister, J, Geyken, A & Kliegl, R 2015, 'childLex: a lexical database of German read by children', *Behavior Research Methods*, vol. 47, no. 4, pp. 1085–94.

Silge, J & Robinson, D 2016, 'tidytext: Text Mining and Analysis Using Tidy Data Principles in R', *JOSS*, vol. 1, no. 3, <http://dx.doi.org/10.21105/joss.00037>.

Soares, AP, Iriarte, Á, de Almeida, José João, Simões, A, Costa, A, Machado, J, França, P, Comesaña, M, Rauber, A, Rato, A & Perea, M 2018, 'Procura-PALavras (P-PAL): A Web-based interface for a new European Portuguese lexical database', *Behavior Research Methods*, vol. 50, no. 4, pp. 1461–81.

Spinner, P 2011, 'Review article: Second language acquisition of Bantu languages: A (mostly) untapped research opportunity', *Second Language Research*, vol. 27, no. 3, pp. 418–30.
Sun, CC, Hendrix, P, Ma, J & Baayen, RH 2018, 'Chinese lexical database (CLD)', *Behavior Research Methods*, vol. 50, no. 6, pp. 2606–29.
Terzopoulos, AR, Duncan, LG, Wilson, Mark A. J., Niolaki, GZ & Masterson, J 2017, 'HelexKids: A word frequency database for Greek and Cypriot primary school children', *Behavior Research Methods*, vol. 49, no. 1, pp. 83–96.
van de Velde, M, Bostoen, K, Nurse, D & Philippson, G (eds) 2019, *The Bantu Languages*, Routledge, New York.
van Heuven, W, Mandera, P, Keuleers, E & Brysbaert, M 2014, 'Subtlex-UK: A new and improved word frequency database for British English', *Quarterly Journal of Experimental Psychology*, vol. 67, no. 6, pp. 1176–90.
Wickham, H 2019, *stringr: Simple, Consistent Wrappers for Common String Operations*, <https://CRAN.R-project.org/package=stringr>.
Wickham, H 2021, *rvest: Easily Harvest (Scrape) Web Pages*, <https://CRAN.R-project.org/package=rvest>.

GiellaLT: an infrastructure for rule-based language technology tool development

Pirinen, Flammie A
UiT Arctic University of Norway
flammie.pirinen@uit.no

Trosterud, Trond
UiT Arctic University of Norway
trond.trosterud@uit.no

Moshagen, Sjur N.
UiT Arctic University of Norway
sjur.n.moshagen@uit.no

Wiechetek, Linda
UiT Arctic University of Norway
linda.wiechetek@uit.no

Abstract

Currently, machine learning is presented as the ultimate solution for language technology regardless of use case and application, however, it requires as a starting point a massive amount of curated linguistic data in electronic form that is expected to be high quality and representative of the kind of language usage that the tools will follow. For minority and indigenous languages, this can be an insurmountable task, as digital materials of the necessary sizes do not exist and can not easily be produced. In this article we present an approach we have successfully used for supporting indigenous languages to survive and grow in digital contexts for years, and describe the potential of our approach for African contexts. Our technological solution is a free and open-source infrastructure that enables language experts and users to cooperate on creating linguistic resources like dictionaries and grammatical descriptions. In addition we provide language-independent frameworks to build these into applications that are needed by the language community.

Keywords: Rule-based LT, keyboards, proofing tools, LT infrastructure, Extremely low-Resource LT

Introduction

Machine learning (ML) approaches have dominated *Natural Language Processing* (NLP) during the last two decades. They typically require a big amount of noise-free data, which does not exist for 99% of the 7,000 world's languages, either due to the small number of writers or a young written norm. Even if a small language community like the Inari Sámi one in Finland is exceptionally productive in text writing (10,533 words/speaker as opposed to 1,440 words/speaker for Swedish) the amount of text recommended for regular machine learning approaches ("A 3.4 billion word text corpus was used for the original BERT-Large, so it is worth training with a data set of this size."[1]) cannot be reached. The same counts for language communities with many speakers but with a low degree of literacy in this language or simply missing written language domains or a bilingual context where the majority language only is used as a written language.

In this article, we present óur infrastructure – *GiellaLT* – presently hosting language models for more than 130 languages and building a number of language technology tools that are useful for any language community. *GiellaLT* is based on an alternative to corpus-based language technology – knowledge-based, also known as rule-based, language technology. The technology is chosen for its ability to support languages with no earlier digital presence. As long as a project has access to a native speaker and a linguist, maybe even in one and the same person, useful tools can be made. Digital resources are always welcome and will help speed up the development work, but they are not a requirement.

A language community with no or little earlier digital presence also needs different types of tools than languages with billion-word corpora. Every language community is unique, but for ones that have or are aiming for a written tradition digitised, typically the first tool to develop is a keyboard, including mobile keyboards nowadays, to be able to enter text correctly and efficiently. No machine learning can create a keyboard layout or discuss language com-

munity needs.

The GiellaLT infrastructure provides all the building blocks to get started on the language work right away, based on open-source technologies and solutions. There is proven integration with most existing systems and platforms, saving huge amounts of time, money, and resources for newcomers. This is especially important for many indigenous language communities, which would not have the resources to develop underlying technologies and integration solutions on their own. By separating language-independent parts from language-specific ones, the cost of developing the language-independent parts can be shared by everyone, or be carried by communities with enough resources. Being open source we are also maintaining an approach that empowers the language communities in a way that retains their ownership of the language and linguistic data they are working on. In the free open-source model, there is no large danger that someone working on the language models simply acquires the linguistic data from the community for free in order to sell it back to them at a higher price.

One final consideration that our rule-based approach has over the machine-learnt models is one of efficiency: a neural language model has to be trained on a system with at least one GPU for over a period of days or weeks and should probably also run on similar hardware, or accessed over high-speed internet. The rule-based language tools can be compiled on a low-end home computer and run locally on even lower-end mobile phones, which is almost a necessity for many writers' tools. As an added bonus the approach is energy-efficient which is something neural models at the moment still struggle with, c.f. Treviso et al. (n.d.).

GiellaLT – A multilingual infrastructure for everyone

The foundation for the work presented in this article is the multilingual infrastructure *GiellaLT*, which includes numerous languages that have little or no data, a rare case in the NLP world. Everything produced in the *GiellaLT* infrastructure is under free and open licences and freely available. The corpora are available with free licensing where possible. The infrastructure is split code-wise into three GitHub organisations: *GiellaLT* containing the language data for each language, *Divvun* containing language-independent code for the infrastructure, and *Giellatekno* for corpus infrastructure. End user tools served by the Divvun group are at *divvun.no* & *divvun.org*, and tools served by the Giellatekno group at *giellatekno.uit.no*, both at *UiT—Norway's Arctic University*.

The basic requirement for developing language tools in GiellaLT is an orthography that is either defined beforehand or is being defined in making the language tools. Access to a printed dictionary is of great help, even more so if it is available electronically. An existing grammatical description is also of tremendous help, but not a requirement.

We build systems that include lexical data as well as rules governing morphophonology, syntax, and semantics as well as a number of application-specific information, e.g. grammatical rules for grammar checking, phonetic rules for *Text-To-Speech* (TTS), and so forth.

The language-dependent work is done within the infrastructure in language-specific repositories, the language-independent features and updates that are relevant to all languages are semi-automatically merged as they are developed. To ensure that language-independent and common features and updates do not destroy existing language data or degrade the language tools, we enforce a rigorous continuous integration-based testing regime. The current system for testing is a combination of our long-term investment in testing within the infrastructure locally for developers—combined with modern automatic testing currently supplied by GitHub actions.

Another part of the *GiellaLT* philosophy is that of reusable and multi-purposeful resources, cf. Antonsen et al. (2010). This is true for all of our work, from corpus collection to cross-lingual cooperation and is crucial for the sustainability of the work in indigenous and lower-resourced languages

where language experts' work time is scarce and precious.

Despite the lack of data, there are high-level tools in *GiellaLT* such as *machine translation* (MT), text-to-speech *TTS*, spelling and grammar checkers, and more, that have been very well received in the language communities. This would not have been possible without first developing basic tools such as keyboards, morphological analysers, and spelling checkers.

GiellaLT for African languages

In this section, we go through the NLP tools GiellaLT infrastructure provides for language users. We list here a subset of the applications that we have found are most useful for the language communities we work with, in language support and revitalisation work: Keyboards and spelling and grammar checking and correction, dictionaries and machine translation. We also provide tools for speech technology, however, this is not discussed in detail in this article, for example, c.f. Hiovain-Asikainen & Moshagen (2022). Furthermore, the linguistic resources that are used as a basis of end-user tools like morphological analysers, are a key resource for digital humanities work on the language: tokenisation, morphosyntactic analysis and glossing for example.

Languages for which Microsoft and Google do not make language technology solutions are vulnerable to technological changes. Closed source programs for such languages run the risk of becoming unusable as word processors or operative system change. Companies behand these programs may then either go bankrupt or change their focus to other areas. Being closed source, the work behind these solutions is then lost. The GiellaTL solution to this is to keep both the linguistic software and the software needed for integrating it in various applications as open source. This means that scarce resources may be reused, without the risk of loosing work. Open access to language independent software also makes it easier to build solutions for lesser-used languages.

Keyboards

Most African languages are written with the Latin alphabet. Many of them have letters outside the A-Z range and even more, have extensive systems of diacritical symbols. On top of that comes the challenge of how click sounds are treated in the San languages, with letter symbols resembling punctuation marks.

Each of these languages does need its own keyboard setup. When language technology tools to an increasing extent are linked to keyboard setups, languages using only the letters A-Z will need their own keyboard to invoke language technology tools for the appropriate language.

In order to meet this challenge, the GiellaLT infrastructure comes with a pipeline for making keyboards and installing them and their corresponding language technology tools on different platforms. The core of the pipeline is the `kbdgen` tool, with which one can easily specify a keyboard layout in a *YAML* file, mimicking the actual layout of the keyboard. The listing below shows the definition of the Android mobile keyboard layout for Lule Sámi. The tool takes this definition and a bit of metadata combines it with code for an Android keyboard app, compiles everything, signs the built artefact and uploads it to the Google Play Store, ready for testing.

```
modes:
    android:
      default: |
            á w e r t y u i o p å
            a s d f g h j k l ø æ
              z x c v b n m ŋ
```

`kbdgen` supports generating keyboard apps or installer packages for Android, iOS, macOS, Windows, Linux (X11 and m17n) and Chrome OS. There is experimental support for generating *Common Language Data Repository* (CLDR) XML

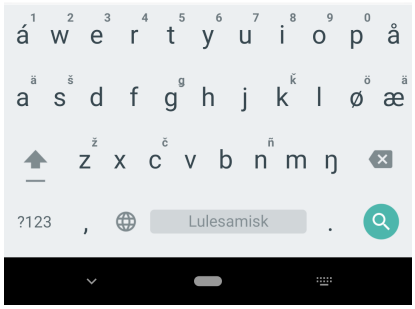

Figure 1: Screenshot of Lule Sámi keyboard for Android, as defined in the listing above.

files, *Scalable Vector Graphics* (SVG) files for fast layout debugging, and finite state transducers for neighbour key mistyping error models. The Windows installer includes a tool to register unknown languages, so that even languages never seen on a Windows computer will be properly registered, thus making it ready to support proofing tools and other language processing tools.

The mobile keyboard app also includes support for spellers (see later), to support the writing process.

Morphological analysers and dictionaries

Not counting Swahili and a few other languages, African languages have very limited corpus resources. Many African languages also have complex morphological structures. This holds most notably for Khoisan and (most) Bantu languages in the south and east as well as Afro-Asiatic languages in the northeast. Such languages, having complex morphology and little or no text resources, will be problematic for mainstream language technology.

Not much lnguage technology has been developed for these languages. The main exceptions are the works by Arvi Hurskainen on Swahili (for an overview and references, see Hurskainen (2018)) and the works by Laurette Pretorius, Sonja Bosch and others for Zulu, Xhosa and other South African languages (e.g. Pretorius & Bosch (2009*a*), Pretorius & Bosch (2009*b*)).

The GiellaLT infrastructure was developed to deal with this situation. The challenge of making language technology for complex and marginalised circumpolar languages was solved by representing the morphological structure as *finite state transducers*. Morphological analysers are the core of our language technology tools, they are written in form of *Finite State Morphology* Beesley & Karttunen (2003), Lindén et al. (2013). This means in practice that language modelling is based on dictionaries and hand-written rules for morphotactics as well as syntactic and semantic processing as needed, e.g. with *Constraint Grammar* Karlsson (1990), Didriksen (2010). While writing rule-based models for language processing is contemporarily written off as too slow and labour-intensive, in our experience full-time work on the dictionary and morphotactic rules for three months is enough to create high-coverage usable language models. If one compares this to the work it takes to create and curate corpora for machine learning, three months does not get one very far in the creation of gigaword corpora.

The main language families treated in the *GiellaLT* infrastructure are Uralic, Algonquin, Eskimo-Aleut and various Siberian languages. As for African languages, we have so far only experimented with eight of them[2]. The Somali language model is far beyond alpha level, it contains almost 16000 stems and the core morphology. The other language models are all relatively small, but in some cases, they still give an impression of how central morphophonological challenges may be solved.

Proofing tools

Proofing tools are a crucial piece of software to support normative language writing, i.e. spelling and grammar checkers and correctors. The morphological analysers described above are the basis for about every other tool one can build using the GiellaLT infrastructure, including proofing tools.

Spellers

A speller consists of essentially two parts: a morphological dictionary that contains the information as to whether a given word is part of the language or not. It is assumed by spell-checking that words not in the collected dictionary are misspellings of real words. The second part is an error model that takes the unknown word user inputted and tries to find similar words that are found in the language.

Technologies similar to the ones presented hare have been applied for African languages as well. One example is Bosch & Eiselen (2012), for Zulu, another is the Swahili speller distributed by the Finnish company Lingsoft.

In the GiellaLT infrastructure, both parts are modelled as finite state transducers (FSTs). The morphologically aware dictionary is built directly on the morphological analyser mentioned above after removing unwanted content such as punctuation and non-standard language. The error model is built as a *Levenshtein* edit distance model Levenshtein (1965) plus language-specific weighted replace rules Pirinen & Lindén (2010).

The speller package is distributed through our own desktop installation and updates system *Páhkat*, and a `pahkat` client is also part of the mobile keyboard apps. This means that spellers on all supported systems, both mobile and desktop, are automatically kept up to date.

As part of the mobile keyboard app, spellers help people in the writing process when using their native keyboard. We are currently experimenting with various forms of word completion and prediction models, but nothing has been released yet.

Grammar checking

The GiellaLT infrastructure includes an advanced grammar checker framework. It uses a combination of morphological analysers and Constraint Grammar Didriksen (2010) disambiguation and error detection components. Furthermore, the constraint grammar logic is used to determine where the grammar errors are; the logic is similar to grammar-based syntax parsing Wiechetek (2012). It is all rule-based, which means that it is possible to develop with essentially no pre-existing electronic corpora. The grammar checker features are quite new but are already used for four different languages.

Machine Translation

The GiellaLT infrastructure supports developing machine translation systems in cooperation with Apertium Khanna et al. (2021). The monolingual models developed in the GiellaLT infra are then combined with the transfer rules and lexicons in Apertium to provide an end-to-end MT system. The Apertium system at its core is also a rule-based machine translation toolkit. This means that one can build MT systems for languages with next to no existing resources in bilingual corpora as well, extending the work put in the monolingual dictionaries. The components needed for a rule-based machine translation on top of the rule-based morphological analysis in GiellaLT infra are a bilingual dictionary, i.e. a regular word-to-word translation dictionary and a set of grammatical rules concerning the translation of linguistic differences between languages.

Conclusion

We have presented the GiellaLT infrastructure and its philosophy and goal of supporting indigenous languages around the world, especially languages with complex morphology or phonology – or both. We have shown that a broad range of useful tools for language communities can be built with minimal preexisting electronic resources, thus allowing the creation of language technology tools for any language, and we have stressed the importance of open source as a strategy for avoiding losing language resources. Finally, we have given an overview of existing resources for African languages in the GiellaLT infrastructure.

The current status of African languages within our infrastructure is limited to several startups and experimental languages, one of the aims of this article is to further survey the potential for future coopera-

tion and engagement in the development of African NLP within our infrastructure.

Notes

[1] Hajdu Róbert 2021: Train BERT-Large in your own language. Towards Data Science.

[2] These are Akan, Amharic, Luo, Ndolo, Pedi, Somali, Tigrinya, and Zulu. The source code is available at *https://github.com/giellalt/lang-xxx*, where *xxx* should be replaced with the ISO 639-3 code of the language in question.

Acknowledgements

All work by the Divvun development group at UiT is funded by the Norwegian Ministry of Local Government and Regional Development. All work by the Giellatekno research group is funded by UiT The Arctic University of Norway.

References

Antonsen, L., Trosterud, T. & Wiechetek, L. (2010), Reusing grammatical resources for new languages, *in* 'Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)', European Language Resources Association (ELRA), Valletta, Malta.
URL: *http://www.lrec-conf.org/proceedings/lrec2010/pdf/254_Paper.pdf*

Beesley, K. R. & Karttunen, L. (2003), 'Finite-state morphology: Xerox tools and techniques', *CSLI, Stanford* .

Bosch, S. & Eiselen, R. (2012), 'The effectiveness of morphological rules for an isiZulu spelling checker', *South African Journal of African Languages* **25**, 25–36.

Didriksen, T. (2010), *Constraint Grammar Manual: 3rd version of the CG formalism variant*, GrammarSoft ApS, Denmark.
URL: *http://visl.sdu.dk/cg3/vislcg3.pdf (Accessed 2017-11-29)*

Hiovain-Asikainen, K. & Moshagen, S. (2022), Building open-source speech technology for low-resource minority languages with sámi as an example – tools, methods and experiments, *in* 'Proceedings of the the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages', European Language Resources Association, Marseille, France, pp. 169–175.
URL: *https://aclanthology.org/2022.sigul-1.22*

Hurskainen, A. (2018), Sustainable language technology for african languages, *in* A. Agwuele

& A. Bodomo, eds, 'The Routledge Handbook of African Linguistics', Routledge Handbooks, Routledge, Abingdon, pp. 359–375.
URL: *https://doi.org/10.4324/9781315392981-19*

Karlsson, F. (1990), Constraint grammar as a framework for parsing unrestricted text, *in* H. Karlgren, ed., 'Proceedings of the 13th International Conference of Computational Linguistics', Vol. 3, Helsinki, pp. 168–173.

Khanna, T., Washington, J. N., Tyers, F. M., Bayatlı, S., Swanson, D. G., Pirinen, T. A., Tang, I. & Alòs i Font, H. (2021), 'Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages', *Machine Translation* pp. 1–28.

Levenshtein, V. I. (1965), 'Двоичные коды с исправлением выпадений, вставок и замещений символов', Доклады Академий Наук СССР **163**(4), 845–848.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T. A. & Silfverberg, M. (2013), Hfst—a system for creating nlp tools, *in* 'International workshop on systems and frameworks for computational morphology', Springer, pp. 53–71.

Pirinen, T. & Lindén, K. (2010), 'Finite-state spell-checking with weighted language and error models: Building and evaluating spell-checkers with wikipedia as corpus', *Proceedings of LREC 2010* .

Pretorius, L. & Bosch, S. (2009*a*), 'Computational aids for Zulu natural language processing', *Southern African Linguistics and Applied Language Studies* **21**, 267–282.

Pretorius, L. & Bosch, S. (2009*b*), Exploiting cross-linguistic similarities in Zulu and Xhosa computational morphology, *in* 'Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages', Association for Computational Linguistics, p. 96–103.

Treviso, M., Ji, T., Lee, J.-U., van Aken, B., Cao, Q., Ciosici, M. R., Hassid, M., Heafield, K., Hooker, S., Martins, P. H., Martins, A. F. T., Milder, P., Raffel, C., Simpson, E., Slonim, N., Balasubramanian, N., Derczynski, L. & Schwartz, R. (n.d.).
URL: *https://arxiv.org/abs/2209.00099*

Wiechetek, L. (2012), Constraint Grammar based correction of grammatical errors for North Sámi, *in* G. D. Pauw, G.-M. de Schryver, M. Forcada, K. Sarasola, F. Tyers & P. Wagacha, eds, 'Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL 8/AFLAT 2012)', European Language Resources Association (ELRA), Istanbul, Turkey, pp. 35–40.

The Analysis of the Sepedi-English Code-switched Radio News Corpus

Ramalepe, Simon
Computer Science Department, University of Limpopo, Sovenga, 0727, South Africa.
simon.ramalepe@ul.ac.za

Modipa, Thipe I.
Computer Science Department, University of Limpopo, Sovenga, 0727, South Africa.
Centre for Artificial Intelligence Research (CAIR), South Africa.
thipe.modipa@ul.ac.za

Davel, Marelie H.
Faculty of Engineering, North-West University, Potchefstroom, South Africa
Centre for Artificial Intelligence Research (CAIR), South Africa.

Abstract

Code-switching is a phenomenon that occurs mostly in multilingual countries where multilingual speakers often switch between languages in their conversations. The unavailability of large-scale code-switched corpora hampers the development and training of language models for the generation of code-switched text. In this study, we explore the initial phase of collecting and creating Sepedi-English code-switched corpus for generating synthetic news. Radio news and the frequency of code-switching on read news were considered and analysed. We developed and trained a Transformer-based language model using the collected code-switched dataset. We observed that the frequency of code-switched data in the dataset was very low at 1.1%. We complemented our dataset with the news headlines dataset to create a new dataset. Although the frequency was still low, the model obtained the optimal loss rate of 2,361 with an accuracy of 66%.

Keywords: Code-switching, text generation, radio news, Transformers, Sepedi

1 Introduction

Code-switching is the use of more than one language within a sentence in a discourse. It is generally more prevalent in multilingual communities through speech than text (Gupta et al. 2020). However, the current trend of communication through technology involves text, and it is more eminent on social media. Like in most multilingual countries, South Africans, notably Sepedi speakers use code-switching in their conversations. Code-switching is generally categorised into inter-sentential and intra-sentential switching. Inter-sentential code-switching occurs on a sentence level when a speaker switches languages from one sentence to another while intra-sentential occurs on a word level when more than one language is used in a sentence (Hamed et al. 2017). The primary language in a code-switching environment is commonly known as the matrix language while secondary language is known as the embedded language (Hamed et al. 2017).

Another form of intra-sentential code-switching uses borrowed or loanwords (Chan et al. 2005). Borrowed or loaned words occur in situations where vowels or consonants are either added or replaced to reproduce phonetically accepted words in the matrix language. The distinction between code-switching, code-mixing, and borrowing is difficult hence, in this study we will use code-switching to refer to both inter and intra-sentential code-switching.

Studies have shown that there is sparsity or lack of code-switched text data (Chang et al. 2018, Gao et al. 2019, Gupta et al. 2020). This could be a result of the informal nature of code-switched data. In other words, code-switching normally occurs on an informal or social environment either in speech or text. Hence, there is a lack of documented datasets on code-switched data. Development of synthetically generated code-switched corpus that can be used to train language models for text generation has been proposed. However, the generated text is not au-

thentic and has to be tested if it resembles real code-switched phenomenon. The lack of existing large-scale code-switched corpora is a challenge to the development of code-switched text generation language models, especially for under-resourced languages such as Sepedi.

Sepedi is one of the official languages of South Africa and is classified as an under-resourced language. Currently, the available code-switched corpora for the Sepedi-English language pair is not enough for use with deep learning techniques. In this study, we discuss and analyse the initial process of collecting Sepedi-English code-switched text data for the development of code-switched models for this language pair. The collected corpus is then applied to a text generation language model to observe the performance of the model using non synthetically generated code-switched text data.

The paper is structured as follows: Section 2 discusses the background of the study. Section 3 discusses the datasets used for training the text generation model. In Section 4 we discuss the experiments conducted, while in Section 5 we focus on the results obtained and the evaluation of the model thereof. Concluding remarks and future work are made in Section 6.

2 Background

Sepedi language is mainly spoken in the Limpopo province of South Africa with a population of 5.9 million (Statistics South Africa 2022). It is the matrix language in most spontaneous conversations in the province while English and other South African languages become embedded in the conversations. Code-switching is not common in read speech or written text but occurs quite often in spontaneous conversations which makes it difficult to collect code-switched data for training and testing language models. When it does occur it is for emphasis in a multilingual setting. Attempts to collect code-switched speech corpora have been made. Hamed et al. (2018), collected spontaneous Egyptian Arabic-English code-switched speech data by conducting and recording informal interview conversations. Analysis of their data showed that there was high usage of code-mixing (intra-sentential). Of the 1,234 sentences in the their dataset, 985 (79.8%) sentences were code-mixed, 124 (10%) sentences were Arabic monolingual text and 125 (10.1%) sentences were English monolingual text. The most frequently trigger words preceding the code-switching point were also noted along with the most frequent uni-grams, bi-grams, and tri-grams. Part-of-speech (POS) tagging was done to a portion of data and it was noted that nouns are used mostly in the embedded language.

In another study, Chan et al. (2005) developed a Cantonese-English code-mixing speech corpus to study the effect of Cantonese accent in English. A situation where most of the English words contain a Cantonese accent. For data collection, newsgroups and online diary methods were used. The frequency of code-switched words, part-of-speech tagging of the code-mixed words, and the length of the code-switched text were noted. Like in Hamed et al. (2017), the frequency of noun occurrence was high at 62.3% and words with length 1 were 74.96% Lyu et al. (2015) also, developed a Mandarin-English Code-switching Speech Corpus. Their corpus was dominated by intra-sentential code-switching of the matrix language. The data collection approach also included recorded interviews and conversations. The most frequent words were also identified.

Modipa et al. (2013), developed a Sepedi-English code-switched speech corpus to analyse the implication factors of code-switching when developing an automatic speech recognition (ASR) systems that are capable of dealing with Sepedi-English code-switched speech. In their data collection approach, radio broadcasts were recorded and the number of code-switched events was counted and transcribed to create the Sepedi prompted code-switched corpus (SPCS) (Modipa & Davel 2022).

Table 1: The number of sentences, words and unique tokens in the datasets.

Dataset	Categories	Total	Train/Val/Test
Radio News	#Sentences	501	350/100.1/50
	#Tokens	14 516	
	#Unique tokens	1 654	
	%English words	1.1%	
News Headlines	#Sentences	1 182	827/236/119
	#Tokens	16 135	
	#Unique tokens	2 781	
	% English words	25%	
Combined dataset	#Sentences	1 683	1 178/337/168
	#Tokens	30 654	
	#Unique tokens	3 824	
	% English words	26%	

Table 2: Top 10 most frequent English and borrowed words in the dataset.

English Word	**Occurrence**	**Borrowed Word**	**Occurrence**
National	11	praevete	1
Economic	12	Magistrata	1
Andrew	13	yaSouth	1
Congress	14	Peresente	1
Africa	14	Uniti	1
Eskom	15	Ekonomi	1
Democratic	17	Dimillione	4
African	20	konferenseng	12
Clip	25	Unione	18
Cyril	28	probenseng	46

Although the SPCS is small (contains short code-switched phrases), it does provide a baseline for code-switched corpus for the Sepedi-English language pair. Marivate et al. (2020) created Sepedi news headlines corpus from a national radio news' Facebook page. The data was used to develop a news classification model.

Van Der Westhuizen & Niesler (2016) compiled the spontaneous English-isiZulu code-switched speech corpus from the South African soap operas. The data was manually transcribed and monolingual English text dominated the corpus by 75%. The data was annotated and code-switching boundaries were identified. Multilingual code-switched corpus for English-isiZulu, English-isiXhosa, English-Setswana, and English-Sesotho have been developed (Van Der Westhuizen & Niesler 2018). Data collection for this multilingual code-switched corpus was obtained from digital video recordings of 626 South African soap opera episodes. The data was manually transcribed by the fluent bilingual speakers of the language pairs. The most code-switching trigger words were identified in each of the language pairs.

In another study, Jansen van Vueren & Niesler (2021) used data augmentation to train and evaluate the performance of code-switched language models. Long short-term memory (LSTM) was used as a generative model to synthetically generate code-

switched data to augment the small code-switched datasets. The study observes that optimised models could generate text with an improved perplexity and word-error rate as compared to models without data optimisation that were studied in (Van Der Westhuizen & Niesler 2016).

3 Data collection

The Sepedi radio news and News headlines datasets are used in this study for the development of the models and analisis. The Sepedi radio news dataset is the primary dataset for this study. The data was collected from a community radio station based in Limpopo to create a code-switched corpus. Several broadcast shows are presented daily, and the data collection focused on radio news read during the various times of the day between June and August 2022. Data cleaning was performed to standardise the data. We used the dataset to train the developed Transformer-based text generation model and observed its accuracy when generating synthetic news.

Table 1, shows the size of the datasets with a split of 70% for training, 20% for validation, and 10% for testing". The news headlines dataset Marivate et al. (2020) was used for comparison with the Sepedi radio news dataset. The dataset is relatively small with 1 182 sentences with minimal code-switching.

Table 2, shows the top 10 English and borrowed words in the Sepedi Radio news dataset. The total number of English unique tokens is 136 which constitutes just 8.8% of the total unique tokens in the dataset. Only 9% of English unique tokens have a frequency of 10 and higher. The average number of words per sentence in the combined dataset is 18.2. We randomly sampled 30 sentences in the radio news dataset and observed that English nouns and proper nouns dominated the form of code-switching in the dataset along with the usage of borrowed words as can be seen in Table 2. The number of English words were 31 out of 929 words in the sampled data. This translates to a low frequency of code-switching of just 3.3% with a ratio

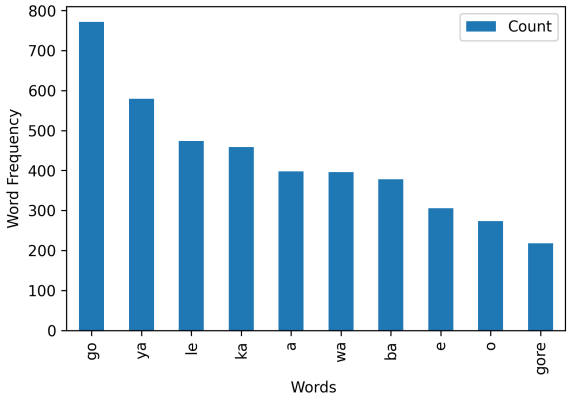

Figure 1: Word frequency in the training dataset

of just 1 English word per sentence. Further analysis of the dataset shows no inter-sentential code-switching.

Data visualisation in all datasets shows that bigrams (character level) have a high frequency. Fig. 1 shows the top 10 word frequencies in the Sepedi radio news dataset. The y-axis depicts the frequency of each word in the dataset while the x-axis shows the words in the vocabulary sorted from most to least frequent.

4 Experimental Setup

For comparison with the previous study by Ramalepe et al. (2022) on monolingual data we adopted the same approach that they used to develop the Transformer based model. The developed model has one Transformer block with causal masking on the attention layers, two separate embedding layers for tokens and a token index with one dense layer with 2 attention heads. We used 64 embedding size for model complexity with the default dropout rate of 0.1. Adam was used as the model optimiser and the rectified linear unit (ReLU) as the activation function. The vocabulary size was set 3k (almost the total number of unique tokens in the combined dataset). Both datasets are combined to observe if there is an improvement in the model's performance as the size of the data increases. The model was trained for 50 epochs due to the small amount of data in the dataset.

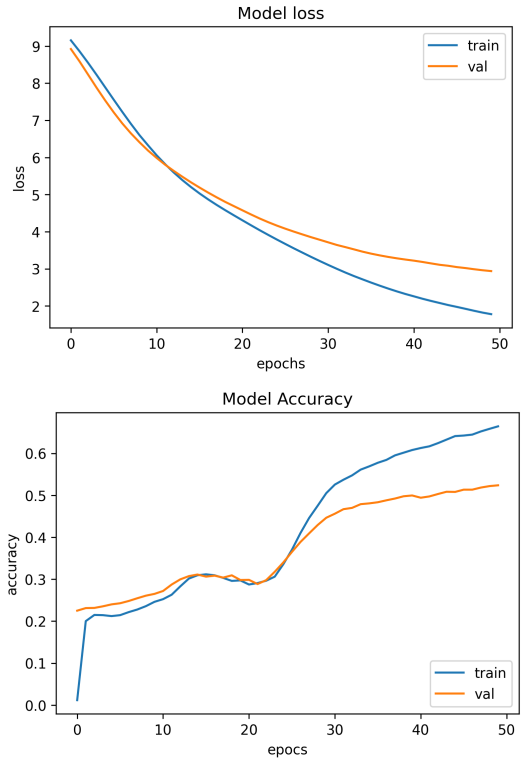

Figure 2: Loss and accuracy curve for the Radio news dataset

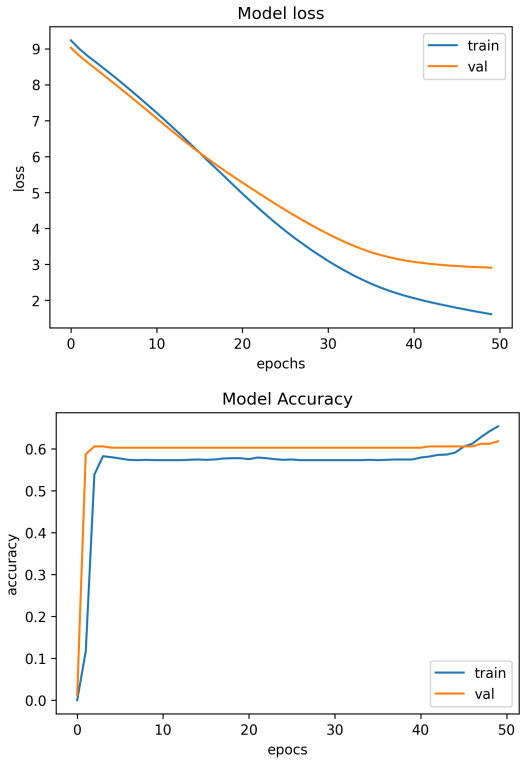

Figure 3: Loss and accuracy curve for the News headlines dataset

The accuracy of the model was measured by computing the total number of correct predictions as a percentage of the total number of predictions during training using the *SparseCategoricalAccuracy* function metric which is used mostly when making text predictions for sparse targets in deep learning. *SparseCategoricalAccuracy* metric checks if the maximum true value is equal to the index of the maximum predicted value.

5 Results and Analysis

Fig. 2 shows the loss and the accuracy curve at each epoch of the training process for the Sepedi radio news dataset, while Fig. 3 shows the loss and the accuracy curve of the news headlines dataset. The loss and accuracy curve for the combined dataset is shown in Fig. 4. The optimal loss rate of the model was 2.361 obtained with the combined dataset with an accuracy of 66%. We observed that although the obtained optimal accuracy was an improvement from 50.3% obtained in Moila & Modipa (2020) using the LSTM based technique on monolingual data, it was still less compared to 75% accuracy obtained in Ramalepe et al. (2022) using the Transformer based approach. Table 3 shows the summary of the results. It is further observed that the size of the dataset influences the accuracy of the model. When the model was trained with the Sepedi radio news dataset(the smallest dataset), the accuracy obtained was low, it increased as we increased the data in the dataset.

In all the figures, Fig. 2, Fig. 3, and Fig. 4 the validation set struggled to generalise, showing an indication of overfitting. This is largely due to the limited amount of data we had in all the datasets. To generate text we start by feeding the model with the starting prompt, the model then generates the conditional probability distribution over the input

Table 3: Validation loss rate and accuracy for each dataset

Name-Dataset	Val-loss error rate	Val-accuracy rate
Radio news	2.938	0.521
News Headlines	3.321	0.61
Combined-datasets	2.361	0.669

Table 4: Generated text

magareng ga tseo di kago letelwa iring ya 13hoo go thobela fm go akaretswa tsa polao ya modiragatsi sibusiso khwinana ba sola anc yo a vandata zinc tša gore magato a gauteng are o mongwe wa marematlou *tsa selegae thekgo hlatlositse* ntwa ya profense ya zululand lehono sa folaga ya matlakadibeseleteng sa go se okobetse ga *pharela ya dienywa tsa citrus go hlola dibaka tsa selegae* go tia ya lehono bodikela mamelodi

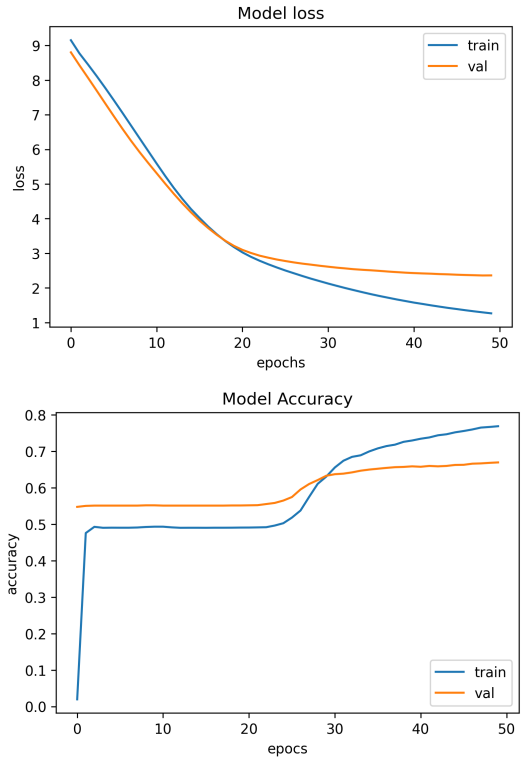

Figure 4: loss and accuracy curve: Combined dataset.

words and sample the next word from the conditional probabilities. The new set of words then becomes the new input to the model, the process continues until the maximum sequence length is reached. To observe the performance of the model on code-switched data we supplied the model with an input text and the model generated the text in Table 4. From the generated text, it can be seen that although the sentences are grammatically correct some are not semantically correct and do not have proper word organisation (formatted in italics in the generated text). For example, the phrase *"tsa selegae thekgo hlatlositse " (of local support increase)* could be corrected as *"hlatlositse thekgo ya tsa selegae"* (increased local support). The number of English words in all the datasets was very small, hence those words did not influence the accuracy of the model and the generated text. However, the performance of the model is low by 12% from the 75% obtained in Ramalepe et al. (2022)

Although the frequency of code-switched words in the datasets was very low, the model could at least generate one code-switched word. For example *"pharela ya dienywa tsa citrus go hlola dibaka tsa selegae"* (The impasse of citrus fruits to create local opportunities). The results signify a positive sentiment of success towards the creation of a large-scale code-switched dataset to train larger models. One major challenge still to be looked at is finding feasible ways of obtaining spontaneous code-switched data as compared to read text. However, such data is often taxing due to the transcription of large amount of speech text.

6 Conclusion

In this study, we discussed the initial phase of collecting, developing, and analysing code-switched corpus using the Sepedi radio news. The developed Sepedi radio news corpus was used to train a Transformer-based text generation model to observe its performance on code-switched data. The model achieved the optimal accuracy of 66% with combined dataset. Data augmentation may be considered in the future to augment the text and create a new code-switched dataset. Other means of collecting spontaneous code-switched data through live recordings and transcription may also be considered in future. To validate the quality of the generated text, human evaluators may also be considered as part of future work.

Acknowledgements

This work is supported by the Centre for Artificial Intelligence Research and the Telkom Centre of Excellence at the University of Limpopo.

References

Chan, J. Y. C., Ching, P. C. & Lee, T. (2005), Development of a Cantonese-English Code-mixing Speech Corpus, *in* 'Ninth European conference on speech communication and technology.'.

Chang, C.-T., Chuang, S.-P. & Lee, H.-Y. (2018), Code-switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation, *in* 'arXiv preprint arXiv:1811.02356.'.

Gao, Y., Feng, J., Liu, Y., Hou, L., Pan, X. & Ma, Y. (2019), Code-switching sentence generation by BERT and generative adversarial networks, *in* 'In INTERSPEECH', Vol. 2019-September, International Speech Communication Association, pp. 3525–3529.

Gupta, D., Ekbal, A. & Bhattacharyya, P. (2020), Findings of the Association for Computational Linguistics A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning, *in* 'In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:', pp. 2267–2280.

Hamed, I., Elmahdy, M. & Abdennadher, S. (2017), Building a First Language Model for Code-switch Arabic-English, *in* 'Procedia Computer Science', Vol. 117, Elsevier B.V., pp. 208–216.

Hamed, I., Elmahdy, M., Abdennadher, S., Tagamoa, E. & Khames, E. (2018), Collection and Analysis of Code-switch Egyptian Arabic-English Speech Corpus, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).', pp. 3805–3809.

Jansen van Vueren, J. & Niesler, T. (2021), Optimised Code-Switched Language Model Data Augmentation in Four Under-Resourced South African Languages, *in* 'Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)', Vol. 12997 LNAI, Springer Science and Business Media Deutschland GmbH, pp. 303–316.

Lyu, D. C., Tan, T. P., Chng, E. S. & Li, H. (2015), 'Mandarin–English code-switching speech corpus in South-East Asia: SEAME', *Language Resources and Evaluation* **49**(3), 581–600.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. (2020), 'Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi'.

Modipa, T. I. & Davel, M. H. (2022), 'Two sepedi-english code-switched speech corpora', *Language Resources and Evaluation* **56**(3), 703–727.

Modipa, T. I., Davel, M. H. & de Wet, F. (2013), Implications of Sepedi/English code switching for ASR systems, *in* 'In Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2013)', Johannesburg, South Africa.

Moila, M. M. & Modipa, T. I. (2020), The development of a sepedi text generation model using long-short term memory, *in* 'Proceedings of the 2nd International Conference on Intelligent and Innovative Computing Applications', ACM, New York, NY, USA, pp. 1–5.

Ramalepe, P. S., Modipa, T. I. & Davel, H. M. (2022), The Development of a Sepedi Text Generation Model Using Transformers [Paper presentation], *in* 'SATNAC-2022', George, Western Cape, pp. 51–56.
URL: *https://www.satnac.org.za/proceedings*

Statistics South Africa (2022), Mid-year population estimates 2022, Technical report.
URL: *www.statssa.gov.za,info@statssa.gov.za,Tel+27123108911*

Van Der Westhuizen, E. & Niesler, T. (2016), Automatic Speech Recognition of English-isiZulu Code-switched Speech from South African Soap Operas, Technical report.

Van Der Westhuizen, E. & Niesler, T. (2018), A First South African Corpus of Multilingual Code-switched Soap Opera Speech, *in* 'Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).'.

Textual Augmentation Techniques Applied to Low Resource Machine Translation: Case of Swahili

Gitau, Catherine
African Institute for Mathematical Sciences Ghana
cgitau@aimsammi.org

Marivate, Vukosi
Department of Computer Science, University of Pretoria, South Africa
vukosi.marivate@cs.up.ac.za

Abstract

In this work we investigate the impact of applying textual data augmentation tasks to low resource machine translation. There has been recent interest in investigating approaches for training systems for languages with limited resources and one popular approach is the use of data augmentation techniques. Data augmentation aims to increase the quantity of data that is available to train the system. In machine translation, majority of the language pairs around the world are considered low resource because they have little parallel data available and the quality of neural machine translation (NMT) systems depend a lot on the availability of sizable parallel corpora. We study and apply three simple data augmentation techniques popularly used in text classification tasks; synonym replacement, random insertion and contextual data augmentation and compare their performance with baseline neural machine translation for English-Swahili (En-Sw) datasets. We also present results in BLEU, ChrF and Meteor scores. Overall, the contextual data augmentation technique shows some improvements both in the $EN \rightarrow SW$ and $SW \rightarrow EN$ directions. We see that there is potential to use these methods in neural machine translation when more extensive experiments are done with diverse datasets.

Keywords: low-resource, data augmentation, machine translation

1 Introduction

There have been several advancements in machine translation and modern MT systems that can achieve near human-level translation performance on the language pairs that have significantly large parallel training resources. Unfortunately, neural machine translation systems perform very poorly on low-resource language pairs where parallel training data is scarce. Improving translation performance on low-resource language pairs could be very impactful considering that these languages are spoken by a large fraction of the world's population.

According to Lisanza (2021), only about 5-10 million people speak Swahili as their native language, but it is spoken as a second language by around 80 million people in Southeast Africa lingua franca, making it the most widely spoken language of sub-Saharan Africa.

Despite the fact that the language is spoken by millions across the African continent, it accounts for less than 0.1% of the internet whereas 58.4% of the Internet's content is in English according to W3Techs (2020), making it a low-resourced language. Even though Swahili is spoken by so many people, there is little extensive work that has been done to improve translation models built for the language. Data that is needed to produce high quality neural machine translation systems is unavailable resulting in poor translation quality.

In computer vision, data augmentation techniques are used widely to increase the robustness and improve the learning of the objects with very little training examples. In image processing, the trained data is augmented by, for example, horizontally flipping the images, random cropping, tilting etc. Data augmentation has now become a standard technique to train deep neural networks for image processing and it is not very common practice in training networks for natural language processing (NLP) tasks such as machine translation. Applying data augmentation techniques in text is not as straightforward as in computer vision because in computer vision, the label and content of the original im-

age is preserved as for natural language processing (NLP) tasks, there is need to retain the context of the sentence after augmentation. There are however several data augmentation methods that have been proven to improve performance on various NLP tasks such as text classification but it is not common practice to apply these data augmentation methods for tasks such as machine translation.

Neural machine translation (NMT) as presented in the work of Sutskever et al. (2014);Cho et al. (2014) is a sequence to sequence task that uses a bidirectional recurrent neural network known as an encoder to process a source sentence into vectors called the decoder which then predicts words in the target language. To be able to train a model that is able to produce good translations, these networks require a lot of parallel data or sentence pairs with words that are occurring in diverse contexts which is not available in low-resource language pairs therefore making the performance of the models quite low. One of the solutions to this problem is to manually annotate the available monolingual data which is time consuming and expensive or to perform unsupervised data augmentation techniques.

In this work we explore some data augmentation techniques that are widely used to improve text classification tasks and investigate their impact on the performance on low resource neural machine translation models for English-Swahili(En-Sw). The most popular text augmentation techniques applied to text classification tasks consist of four powerful operations; synonym replacement, random insertion, random swap and random deletion. These methods have been shown to improve performance in text classification tasks as shown in Wei & Zou (2019). And to better gauge the effect of our data augmentation methods, we compare the results with a baseline model trained on En-Sw datasets.

In summary, our contributions include:

1. We explore and evaluate on NMT task, three data augmentation techniques currently only being used in text classification tasks; Word2Vec based-augmentation which does synonym replacement in the sentences, TF-IDF-based augmentation to insert words in random positions in the sentence as well as use of Masked Language Model based-augmentation that does contextual data augmentation on the text.
2. We show how these data augmentation techniques can be used in NMT tasks.
3. We also extended the textaugment library Marivate & Sefara (2020) to use Fasttext's pretrained models.
4. We present baseline NMT results in BLEU, Meteor and ChrF scores.

This paper is organised as follows; We look at the work that is been done by past authors on data augmentation for low-resource languages first. We also look at the data augmentation techniques and the approaches we used in our study which is described in Section 3. Section 4 describes the model settings that were considered for every data augmentation approach, Section 5 discusses the experimental results and then we conclude with stating limitations and future work as well as conclude at Sections 6 and 7.

2 Literature Review

Work on machine translation to improve machine translation quality on low resource languages is a widely studied problem. In natural language processing (NLP), data augmentation is a popular technique that is used to increase the size of the training data.

One promising approach is the use of transfer learning Zoph et al. (2016). This method proved that having prior knowledge in translation of a different higher-resource language pair can improve translating a low-resource language. A NMT model is first trained on a large parallel corpus to create the *parent model* and continued to train this model by feeding it with a considerable smaller parallel corpus of a low-resource language resulting into the *child model* which inherits the knowledge from the parent model by reusing its parameters. The parent

and child language pairs shared the same target language(English). The use of data from another language can be seen as a data augmentation task in itself and large improvements have been observed especially when the high-resource language is syntactically similar with the low-resource language Lin et al. (2019).

The work of Sennrich et al. (2016) explores a data augmentation method for machine translation known as back translation where machine translation is used to automatically translate target language monolingual data into source language data to create synthetic parallel data for training and is currently the most commonly used data augmentation technique in machine translation tasks. The quality of the backward system while effective, has been shown to negatively affect the performance of the final NMT Model when the target-side monolingual data is limited. Back translation as a method for performing data augmentation in machine translation could deteriorate the Low resource Language - English(LRL - ENG) translation performance due to the limited size of the training data as shown in Xia et al. (2019).

Xia et al. (2019) augment parallel data through two methods: back-translating from ENG to low resource language (LRL) or high resource language (HRL) and converting the HRL-ENG dataset to a pseudo LRL-ENG dataset. They use an induced bilingual dictionary to inject LRL words into the HRL then further modify these sentences using modified unsupervised machine translation framework. Their method proved to improve translation quality as compared to supervised back-translation baselines however, the method requires access to a HRL that is related to the LRL as well as monolingual LRL.

There are other data augmentation methods which have been used in other NLP tasks such as text classification to improve performance. Wei & Zou (2019) show that simple word replacement using knowledge bases like WordNet Miller (1995) can improve performance of classification tasks. Marivate & Sefara (2020) also observe that Word2Vec-based augmentation is also a viable solution when one does not have access to a knowledge base of synonyms such as the WordNet Miller (1995). Kumar et al. (2020) show that seq2seq pre-trained models can be effectively used for data augmentation and these provide the best performance. These data augmentation methods are currently only being used to improve classification tasks and have not yet been utilized in any neural machine translation task to improve performance. In this work we will be looking at how some of these methods can be used to also improve neural machine translation models where the data is low-resourced. In particular, we will explore three data augmentation methods which include: 1) Word2Vec based-augmentation, 2) Tf-idf based augmentation, 3) Masked Language Model based-augmentation and use the additional data to train the NMT model.

3 Methodology

Our goal is to compare different data augmentation methods that are used in text classification tasks with the aim of identifying whether the methods can be used to improve the baseline NMT score. The results are compared across two different datasets and uses in-domain test sets to demonstrate the generalization capability of the models. These experiments are useful to help other researchers gain insights as they work on building better neural machine translation models for low-resourced languages. First, we describe the data that was used to train the models, then the data augmentation methods that we will be using and finally give details of the experiments we performed to test these methods together with the results obtained.

3.1 Training Data

Small amounts of parallel data are available for Swahili-English. The data was received from the work of Lakew et al. (2020) where they released standardized experimental data and test sets for five different languages(Swahili, Amharic, Tigirinya, Oromo and Somali). They collected all available parallel corpora for those five languages from the

Opus corpus Tiedemann (2012) which consists of a collection of translated texts from the web. For this work, we utilized data that includes JW300 Agić & Vulić (2019) and Tanzil Tiedemann (2012) which provides a collection of Quran translations to compare with the baseline results from the work of Lakew et al. (2020).

Table 1 shows the amount of parallel data that was collected. The data was split into train, dev and test sets as in Lakew et al. (2020). We then segmented the data into subword units using Byte Pair Encoding Sennrich et al. (2016) where we learned $20K$ byte pair encoding tokens.

3.2 Baseline

In this approach the Transformer NMT model is trained using Jw300 and Tanzil data combined then tested on different datasets from two different domains (Jw300 and Tanzil). The model is trained with no modifications throughout with standard preprocessing steps such as tokenization, lowercasing and cleaning. This model in this approach serves as a baseline for comparison.

3.3 Data Augmentation methods

We augmented the data using three types of augmentation methods: Word2Vec-based augmentation(synonym replacement), Tf-idf based augmentation(random insertion) and Masked Language Model(MLM)-based augmentation(context based augmentation). We combined the first two augmentation methods and used the Masked Language Model-based augmentation on its own. The Word2Vec and Tf-idf augmentations were done on the source language such that when training an En-Sw model, we augment the English language and when training a Sw-En model we augment the Swahili language. In Masked language modeling the augmentation was only done on the English language.

3.3.1 Word2vec-based augmentation

Word2vec is an augmentation technique mostly used in classification tasks that uses a word embedding model Mikolov et al. (2013) that is trained on publicly available datasets to find the most similar words for a given input word. We use Word Vectors pre-trained on Common Crawl and Wikipedia on both English and Swahili data using fastText Joulin et al. (2017) a library for text representation and classification. We load the pre-trained fastText model for each language into our algorithm to augment the texts by randomly selecting a word in the text to determine their similar words using cosine similarity as a relative weight to select a similar word that replaces the input word as done in Marivate & Sefara (2020). Our algorithm is as illustrated in Algorithm 1 . It receives a string which is the input data and augments the text into five different augmented texts then we use cosine similarity to select the best sentence that is at least 0.85 closer to the original text. The reason for this is that we'd like to retain the contextual meaning of a sentence even after augmentation. We compare the five different augmented sentences and pick the sentence that has a cosine similarity score that is highest. To prevent duplicated augmentations, we drop the sentences that are 100% similar to the original sentence. This augmentation was done on the source language where the corresponding target language sentences remained constant and unchanged. Examples of the augmented sentences can be seen in Table 2.

3.3.2 Tf-idf based augmentation

We created another set of augmented data that uses Tf-idf Ramos (2003). The concept of Tf-idf is that high frequency words may not be able to provide much information gain in the text. It means that rare words contribute more weights to the model. In this case, words that have low Tf-idf scores are said to be uninformative and thus can be replaced or inserted in text without affecting the ground truth labels of the sentence. Here, the words that are chosen to be inserted at a random position in the sentence are chosen by calculating the Tf-idf scores of

Language Pair	Split	Domain JW300	Tanzil	Total
Sw-En	train	907842	87645	1024717
	dev	5179	3505	8684
	test	5315	3509	8824

Table 1: Data statistics showing number of examples/sentences available across four domains

Method	Sentence
English	
Original	The quick brown fox **jumps** past the lazy dog
Word2Vec + tfidf	The quick brown fox **leaps over retrorsum** the lazy dog
Swahili	
Original	Baba na mama yako ni wazuri sana
Word2Vec + tfidf	**Kizee** baba na mama yako ni **wema waar** sana

Table 2: Table showing example of augmented sentences

words over all the sentences and then taking the lowest ones. We therefore insert a new word at a random position according to the Tf-idf calculation. This was also done on the source language only and the corresponding target language sentences remain unchanged. Tf-idf was applied after performing the Word2Vec based augmentation method. This is illustrated in Algorithm 1.

3.3.3 Masked Language Model (MLM) augmentation

Since the above methods do not consider the context of the sentence, we decided to use Masked Language Modeling(MLM) where we used RoBERTa Liu et al. (2019) a transformer model that is pretrained on a large corpus of English data in a self-supervised fashion. It is used to predict masked words based on the context of the sentence. You can find the algorithm used in Algorithm 2. Taking the sentence, the model randomly masks 15% of the words in the input then runs the entire masked sentence through the model and predicts the masked words which helps the model to learn a bidirectional representation of the sentence. In this work, a sentence is passed through our algorithm which then predicts the masked word creating a new aug-

Algorithm 1 Word2Vec and Tf-idf based augmentation

Input: s: a sentence
Output: $\hat{s}$: augmented sentence

Step 1: get similar words of each word in s :
procedure AUGMENT(s) Augmentation of sentence
    $t \leftarrow$ sentence s tokenized
    $u \leftarrow$ unique words from t
    for w *in*(u) **do**
        $\overrightarrow{w} \leftarrow$ find five similar words for w
    end for
Step 2: replace random words in s with similar words and insert Tf-idf word :
    $n \leftarrow 5$
    for _ *in range*(n) **do**
        $w_i \leftarrow$ randomly select a word from s
        $w_0 \leftarrow$ randomly select one similar word for w_i from $\overrightarrow{w}$
        $\hat{s} \leftarrow$ replace w_i with similar word w_0
        $ss \leftarrow$ insert Tf-idf word in random position in s
        $\hat{ss} \leftarrow$ merge $\hat{s}$ and ss
    end for
    return $\hat{ss}$ Augmented sentence
end procedure

mented sentence. Note that this augmentation method was only done on the English language due to lack of enough resources to train a good MLM for the Swahili language.

Algorithm 2 Masked Language Model augmentation algorithm

Input: s: a sentence
Output: $\hat{s}$: augmented sentence
1: **procedure** MLMAUGMENT(s)
Augmentation of sentence
2: $n \leftarrow$ 15% of words in the sentence
3: **for** _ *in range*(n) **do**
4: $w_i \leftarrow$ randomly select a word from s
5: $\hat{s} \leftarrow$ mask word w_i
6: $\hat{s} \leftarrow$ replace masked word with predicted mask
7: **end for**
8: **return** $\hat{s}$ Augmented sentence
9: **end procedure**

For our experiments we combined the augmented sentences for Word2Vec based-augmentation and Tf-idf based data augmentation producing almost triple the original sentences. The MLM-based augmentation methods produced almost double the original parallel sentences. The total training data that was used is as shown in Table 2.

Method & Language Pair	Total
Word2Vec + tfidf (EN-SW)	2952864
Word2Vec + tfidf (SW-EN)	2774186
MLM (EN-SW)	2048613

Table 3: Data statistics showing total data used for training after augmentation

4 Experiments

This section explains in detail the learning and the model settings that were considered for every data augmentation approach.

4.1 Model Settings

All the models were trained using the transformer architecture of Vaswani et al. (2017) using the open-source machine translation toolkit joeyNMT by Kreutzer et al. (2019). The model parameters were set to 512 hidden units and embedding dimension, 4 layers of self-attentional encoder decoder with 8 heads. The byte pair encoding embedding dimension was set to 256. Adam optimizer is used throughout all experiments with a constant learning rate of 0.0003 and dropout was set at 0.3. All the models were trained on 40 epochs.

4.2 Evaluation Metrics

The models were evaluated using in-domain test sets. The performance of the different approaches was evaluated using different translation evaluation metrics: BLEU Papineni et al. (2002), METEOR Banerjee & Lavie (2005) and chrF Popović (2015). BLEU(Bilingual Evaluation Understudy) is an automatic evaluation metric that is said to have high correlation with human judgements and is used widely as the preferred evaluation metric. METEOR(Metric for Evaluation of Translation with Explicit Ordering) is based on generalized concept of unigram matching between the machine translations and human-produced reference translations unlike BLEU and is calculated by getting the harmonic mean of precision and recall. ChrF is a character n-gram metric, which has shown very good correlations with human judgements especially when translating to morphologically rich languages. The higher the score of these metrics means that the system produces really good translations.

5 Results and Discussion

This section describes the results of the three methods; The baseline (S-NMT), the word2vec-based + tfidf (Word2Vec) augmentation and masked language model augmentation(MLM). Table 4 shows the performance of the different data augmentation methods applied in machine translation. The

BLEU scores for the EN ↔ SW, domain specific best performing results are highlighted for each direction with the bolded scores displaying the overall best scores. We observe that in all the test domains, the models trained with the MLM-augmented data performed better than both the baseline and Word2Vec in most cases. These results are highly related to the fact that the MLM-based augmentations are based on contextual embeddings. The drop in performance in some cases can be due to the fact that the structure of the sentence is not necessarily preserved while doing word or synonym replacement thus making the translation not retain it's original meaning. We can also observe that there is a degradation of performance when translating into the low-resource language for the JW300 test data but for models tested on Tanzil, the degradation occurs mostly when translating into English. The Tanzil training data that was used to train the model was quite low compared to JW300 data which explains the low scores for Tanzil as compared to JW300 data. The Word2Vec + Tf-idf based augmentations do not lead to significant improvements of the baseline model, however, the results show there is potential in using these methods in NMT especially the Masked Language Model for augmentation which proved to perform better than the Word2Vec+Tf-idf model

6 Limitations and Future Work

One of the biggest challenges in machine translation today is learning to translate low-resource language pairs with technical challenges such as learning with limited data or dealing with languages that are distant from each other.

This paper shows that we could potentially use simple data augmentation methods in machine translation. In our experiments, we only augment the source language for the Word2Vec based augmentation method and only augment the English sentences for the MLM based augmentations. In Future work, we plan on exploring augmenting the target side of the parallel data in the Word2Vec-based augmentation and compare performance to the source language augmentation as well as testing the model's ability to generalize by using out-of domain datasets. Another experiment that could be explored is the use of the Word2Vec data augmentation method only without the use of Tf-idf word replacement method as it adds more noise to the sentences. We plan on continuing this research and will make available the algorithms used in this paper at *https://github.com/dsfsi/translate-augmentationhttps://github.com/dsfsi/translate-augmentation*

6.0.1 Computational Considerations

Training time took about 1 hour running one epoch using NVIDIA Tesla V100 GPU on Google Cloud on the augmented texts. Running on Colaboratory took about 5 days to run 40 epochs and with limited time on our hands, there is only so much we could experiment. Running these experiments was quite expensive and there needs to be consideration of budgets as well as time so as to run these MT experiments.

7 Conclusion

In this work we proposed the use of different textual data augmentation tasks in neural machine translation using the low-resourced language Swahili. We also showed how one can perform data augmentation on the low resourced language using pre-trained word vectors and presented baseline results in ChrF and METEOR which have never been presented before. Our investigation shows that although the models trained on the augmented texts did not improve on the baseline model, there is still potential to using these methods in NMT tasks with enough compute and more experiments. We hope that this work will set the stage for further research on applying simple augmentation methods that don't require a lot of computation power in low-resource NMT modelling.

Model	Domain	BLEU		METEOR		ChrF	
		en-sw	sw-en	en-sw	sw-en	en-sw	sw-en
S-NMT(BPE)	JW300	45.30	46.54	66.32	62.21	65.92	**71.30**
	Tanzil	27.48	24.66	50.43	**50.41**	**52.51**	46.69
Word2Vec	JW300	45.23	45.52	65.90	66.56	65.02	61.76
	Tanzil	26.29	**25.80**	49.45	42.43	45.78	40.68
MLM	JW300	**45.32**	**46.98**	**66.56**	**70.55**	**65.94**	69.68
	Tanzil	**29.26**	24.86	**58.31**	49.31	48.23	**47.47**

Table 4: BLEU, ChrF, Meteor scores for Swahili ↔ English directions, domain-specific best performing results are in bold.

8 Bibliographical References

References

Agić, Ž. & Vulić, I. (2019), JW300: A wide-coverage parallel corpus for low-resource languages, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics.
URL: *https://www.aclweb.org/anthology/P19-1310*

Banerjee, S. & Lavie, A. (2005), METEOR: An automatic metric for MT evaluation with improved correlation with human judgments, *in* 'Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization', Association for Computational Linguistics, Ann Arbor, Michigan, pp. 65–72.
URL: *https://www.aclweb.org/anthology/W05-0909*

Cho, K., van Merriënboer, B., Bahdanau, D. & Bengio, Y. (2014), On the properties of neural machine translation: Encoder–decoder approaches, *in* 'Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation', Association for Computational Linguistics, Doha, Qatar, pp. 103–111.
URL: *https://aclanthology.org/W14-4012*

Joulin, A., Grave, E., Bojanowski, P. & Mikolov, T. (2017), Bag of tricks for efficient text classification, *in* 'Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers', Association for Computational Linguistics, Valencia, Spain, pp. 427–431.
URL: *https://aclanthology.org/E17-2068*

Kreutzer, J., Bastings, J. & Riezler, S. (2019), Joey NMT: A minimalist NMT toolkit for novices, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations'.
URL: *https://www.aclweb.org/anthology/D19-3019*

Kumar, V., Choudhary, A. & Cho, E. (2020), Data augmentation using pre-trained transformer models, *in* 'Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems', Association for Computational Linguistics.

Lakew, S. M., Negri, M. & Turchi, M. (2020), 'Low Resource Neural Machine Translation: A Benchmark for Five African Languages', *arXiv e-prints* p. arXiv:2003.14402.

Lin, Y.-H., Chen, C.-Y., Lee, J., Li, Z., Zhang, Y., Xia, M., Rijhwani, S., He, J., Zhang, Z., Ma, X., Anastasopoulos, A., Littell, P. & Neubig, G. (2019), Choosing transfer languages for cross-lingual learning, *in* 'Proceedings of the 57th Annual Meeting of the Association for Compu-

tational Linguistics', Association for Computational Linguistics.

Lisanza, V. (2021), *Swahili gaining popularity globally*. Accessed: 2020–30.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019), 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv e-prints* p. arXiv:1907.11692.

Marivate, V. & Sefara, T. (2020), 'Improving short text classification through global augmentation methods', *Machine Learning and Knowledge Extraction* .
URL: *http://dx.doi.org/10.1007/978-3-030-57321-8_1*

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger, eds, 'Advances in Neural Information Processing Systems', Vol. 26, Curran Associates, Inc.

Miller, G. A. (1995), 'Wordnet: A lexical database for english', *COMMUNICATIONS OF THE ACM* **38**, 39–41.

Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. (2002), Bleu: a method for automatic evaluation of machine translation, *in* 'Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics', Philadelphia, Pennsylvania, USA.
URL: *https://www.aclweb.org/anthology/P02-1040*

Popović, M. (2015), chrF: character n-gram F-score for automatic MT evaluation, *in* 'Proceedings of the Tenth Workshop on Statistical Machine Translation', Lisbon, Portugal.
URL: *https://www.aclweb.org/anthology/W15-3049*

Ramos, J. (2003), Using tf-idf to determine word relevance in document queries.

Sennrich, R., Haddow, B. & Birch, A. (2016), 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)'.
URL: *https://www.aclweb.org/anthology/P16-1162*

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *CoRR* **abs/1409.3215**.
URL: *http://arxiv.org/abs/1409.3215*

Tiedemann, J. (2012), Parallel data, tools and interfaces in OPUS, *in* 'Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)', European Language Resources Association (ELRA).
URL: *http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_paper.pdf*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention Is All You Need', *arXiv e-prints* p. arXiv:1706.03762.

W3Techs (2020), *W3Techs. Usage of content languages for websites*. Accessed: 2020–30.

Wei, J. & Zou, K. (2019), EDA: Easy data augmentation techniques for boosting performance on text classification tasks, *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)', Association for Computational Linguistics, Hong Kong, China, pp. 6382–6388.
URL: *https://www.aclweb.org/anthology/D19-1670*

Xia, M., Kong, X., Anastasopoulos, A. & Neubig, G. (2019), Generalized data augmentation for low-resource translation, *in* 'Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Florence, Italy, pp. 5786–5796.
URL: *https://www.aclweb.org/anthology/P19-1579*

Zoph, B., Yuret, D., May, J. & Knight, K. (2016), Transfer learning for low-resource neural ma-

chine translation, *in* 'Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, Austin, Texas, pp. 1568–1575.
URL: *https://www.aclweb.org/anthology/D16-1163*

Topic modelling to support English text selection for translation into South Africa's other official languages

Mazarura, Jocelyn
Department of Statistics, University of Pretoria
jocelyn.mazarura@up.ac.za

de Wet, Febe
Department of Electrical and Electronic Engineering, Stellenbosch University & School of Electrical, Electronic and Computer Engineering, North-West University
fdw@sun.ac.za

Abstract

Appropriate training data is a prerequisite for the development of natural language processing (NLP) techniques. Vast amounts of language data are typically required to develop NLP tools that perform at state-of-the-art level. Such abundant resources are currently only available in a few languages. The remaining languages have to find alternative ways to become "NLP-enabled". The aim of the study reported on here is to make more language data available to support NLP development in the official languages of South Africa. In this paper we present the idea of generating text data by means of translation. We also propose the use of topic modelling to identify text in a highly resourced source language that will yield meaningful translations in under-resourced target languages. More specifically, the paper describes how topic modelling was used to identify English Wikipedia articles that should be suitable for translation into South Africa's 10 other official languages.

Keywords: DHASA, topic modelling, text selection, translation, under-resourced languages

1 Introduction

The resources required to implement state-of-the art performance in natural language processing (NLP) applications like text generation, question answering, automatic speech recognition, etc. are currently only available in a few well-resourced languages. The other 6 000-odd languages that are spoken in the world today have to find alternative ways to prepare themselves for a digital future society in which the ability to process language and speech automatically will be a prerequisite for participation. Applications like search engines and chat bots are developed using text data while others, like automatic dictation systems, require both text and speech data to be implemented. The current study is specifically aimed at developing text resources that are suitable for language modelling in South Africa's official languages.

A number of text resources have been developed during previous projects and are available in the resource catalogue of the South African Centre for Digital Language Resources (SADiLaR[1]). However, the extent of the existing text corpora is not adequate to implement large vocabulary automatic speech recognition (ASR). Moreover, the most extensive text resource that is available in all the official languages, the NCHLT Text Corpora (Eiselen & Puttkammer 2014), was derived from documents published on the South African government website domain (*.gov.za). As a result, the vocabulary and language usage is domain specific. The aim of the current work is to develop new text resources that can be used in addition to or in combination with existing corpora. The new data should ideally represent a diversity of domains and should contribute to expanding existing lexica.

One obvious and potentially rich source of text data is the world wide web. While many languages have at least some presence on the web, there are also many which are not represented on the internet at all. However, the fact that a language is present on the web does not guarantee that the available data is suitable for NLP research. Web-text is also not by default of good quality. In addition, some internet texts are extremely domain specific. Especially for under-resourced languages, online texts tend to be restricted to religious publications or government

documents.

A strategy that has been proposed to strengthen the internet presence of under-represented languages is to create text by, for instance, encouraging first language speakers to write Wikipedia articles (Pretorius & Wolff 2020). The approach has been successful in some instances and have resulted in the first ever contributions to Wikipedia in many languages. Despite the advantage of producing "native" text, this is a slow process and has, to date, not yielded the amount of text data required by the data-hungry deep learning methods that are currently widely used in the field of NLP.

This paper proposes large scale translation as an alternative to text generation[2]. Despite the differences between *real* and translated text, translations could bridge the gap between having almost no data available at all and having just enough data to bootstrap semi-supervised systems that can contribute to developing further resources. Proposing translation as a means to generate text in an under-resourced language also means deciding on what should be translated. Many of the obvious answers to this question, e.g. the 100 most popular Mandarin Wikipedia articles, would not necessarily yield meaningful text in other languages. The research reported on in this paper represents an attempt to find a systematic way of identifying text that is suitable for translation and that would yield meaningful text in the target languages. More specifically, we describe how English Wikipedia articles that should be suitable for translation into South Africa's 10 other official languages were identified.

2 Background

Previous attempts to translate NLP tasks and data sets from highly resourced to under-resourced languages have encountered various challenges related to the differences between the source and target languages. For instance, during the development of the *African Wordnet*, it was found that many concepts in existing Wordnets were not lexicalised in South African languages (Griesel & Bosch 2020). This difference was overcome by using the SIL Comparative African Wordlist (SILCAWL) introduced by Snider & Roberts (2004) as a guideline to add new synsets to the African Wordnet. This list consists of a vast selection of words relating to various things, such as body parts, emotions and stages of life in English and French. The content of the entries in the list is – to a large extent – lexicalised in South Africa's official languages. Using the SILCAWL as a point of departure has the additional advantage that it does not perpetuate culturally and cognitively biased language resources.

Another example of an NLP resource that was created by translation is the *FLORES* dataset, a multilingual resource that was compiled to enable benchmarking in low-resource and multilingual machine translation (Goyal et al. 2022). The dataset consists of 3001 sentences translated into – at the time of writing – 200 languages, but this number is continuously expanding. Afikaans, Sepedi, SiSwati, Xitsonga, Setswana, isiXhosa and isiZulu are already included in the FLORES set. A translation into Sesotho will be released soon, followed by isiNdebele and Tshivenda.

The FLORES sentences were selected from 842 English Wikipedia articles. Articles were chosen to represent a wide range of topics including crime, disasters, entertainment, geography, health, nature, politics, science, sports, and travel. One might argue that more translations should be generated by expanding the existing list of articles with more articles on the same or similar topics. However, topics were assigned manually to the sentences in the FLORES collection which means that expanding on or repeating the process is not possible. Moreover, further inspection of the FLORES dataset revealed that the source domains, *Wikinews*, *Wikijunior* and *WikiVoyage*, contain many articles that may not always be relevant in the African context and which would probably give rise to the same lexicalistion issues that were encountered during the African Wordnet project.

The *No Language Left Behind Seed Data* (NLLB-Seed) was created in a similar manner as the FLORES data and comprises a set of professionally-

translated Wikipedia sentences (NLLB Team et al. 2022). However, in contrast to the FLORES data, the NLLB-Seed sentences were taken from Wikimedia's so-called "list of articles every Wikipedia should have". The list includes topics in different fields of knowledge and human activity, similar to those in the SILCAWL. The NLLB-Seed data currently consists of around six thousand sentences in 39 languages. Although the list could be a good point of departure for an extensive translation effort, it is not currently foreseen that the list will expand over time. In contrast, the topic modelling method proposed here could generate any number of sentences and could even be repeated if required.

The remainder of this paper is structured as follows. Section 3 provides an overview of topic modelling and Section 4 explains how models are evaluated. Section 5 describes how data was selected to create a Wikipedia topic model related to the SILCAWL. The results of the investigation are presented in Section 6, followed by a discussion in Section 7.

3 Topic modelling

Topic modelling is a text mining technique used to uncover latent topics in large collections of documents. Most topic models are unsupervised, thus making them powerful tools in addressing real-world problems as data is typically unlabelled in practice. Topic models have found use in many applications, including sentiment analysis (Rana et al. 2016), text classification, document categorisation, summarisation, etc. (Boyd-Graber et al. 2017).

Latent Dirichlet Allocation (LDA) is one of the most popular topic models (Blei et al. 2003). It is a three-level hierarchical Bayesian model, in which each document is modelled as a finite mixture over a set of latent topics, where a topic is defined as a distribution over the words of the vocabulary. LDA is a generative probabilistic model as it models the assumption that each document is formed through the following generative process:

1. Assuming that the corpus contains K topics, randomly choose K topic distributions, ϕ_k.
2. For each of the M documents in the corpus, randomly select a distribution over topics, θ_m.
3. Assume that each document contains n_m words. For each word, w_{mn}, in the m^{th} document.
    a. Randomly choose a topic, from the distribution over topics from Step 2, where z_{mn} denotes an indicator of the selected topic.
    b. Randomly choose a word from the selected topic based on its distribution from Step 1.

This generative process can be summarised graphically as in Figure 1.

LDA can be easily applied to any dataset using the gensim[3] package in Python. The main outputs of interest are typically the topics and topic distributions of each document. Gensim also has the capability to identify topic distributions in unseen documents, which is a feature that is useful for classifying new documents.

4 Model evaluation

By virtue of being unsupervised, evaluating topic models is a challenging task. Although some authors use perplexity[4] or held-out likelihood as measures of topic model performance, it has been shown that such intrinsic measures of model performance do not correlate with human understanding of topics (Chang et al. 2009). In other words, models with better perplexity scores often produce less humanly interpretable topics, which defeats the exploratory goals of topic modelling. In light of this, coherence measures have become more popular. In this research, we make use of the well-known UMass coherence measure as it has been shown to align well with human evaluations of coherence (Mimno et al. 2011). Considering the top N words in a topic, the UMass coherence for a topic is calculated in gensim as in Equation 1.

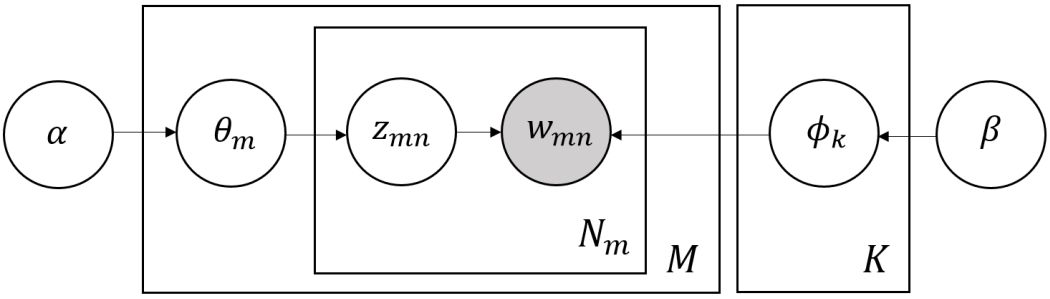

Figure 1: Graphical model for LDA. The shaded circles are used to represent observed variables, whilst unshaded circles represent unobserved variables. The arrows represent dependencies between variables and rectangles indicate replicated structures. α and β denote hyperparameters.

$$C_{\text{UMass}} = \frac{2}{N \cdot (N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \frac{P\left(w_i, w_j\right) + \varepsilon}{P\left(w_j\right)} \tag{1}$$

$P\left(w_j\right)$ denotes the probability of word w_j occurring in the training set of documents. $P\left(w_i, w_j\right)$ denotes the probability of words w_i and w_j occurring together in the training set and ε denotes a smoothing parameter. The overall coherence for a topic model is then taken as the average of all the topic coherences. In general, a higher coherence score is desired as it indicates better topic coherence.

5 Document selection process

The objective of this research is to create a collection of sentences that can be used for NLP research in under-resourced languages, in a manner that is both methodological and systematic. To this end, we propose using topic modelling as a selection method. This section provides the details of the selection process.

5.1 Topic model training

Given the aim of our research, we sought to create a data set that is generally relevant and not domain specific by selecting sentences from Wikipedia articles. To this end, we first created a training set, based on articles from various general topics. These general topics were taken from the SILCAWL comparative word list introduced in Section 2. The English words in the list were used as search terms in Wikipedia and the associated articles were retrieved from the web. After cleaning the data and removing any potentially offensive words, the resulting corpus contained 1 698 documents with an average of 1 132 words per document.

A topic model was then trained on this data using gensim. The hyperparameters were learned from the data, but the model was run for different numbers of topics, $K = \{100, 200, ..., 1000\}$. The average coherence score for each model is shown in Figure 2. It can be seen that the coherence does not increase significantly beyond $K = 800$, thus $K = 800$ was chosen to be the ideal number of topics for the training data.

Table 1 shows the top 25 topics from $K = 800$ arranged in order of decreasing coherence. It is evident from the table that the topic model was able to pick up various topics including, for example, poverty (topic 98), health (topic 224) and time (topic 551). Note that the choice of overall topic names are determined by the user and thus are subjective. The full set of topics can be found at https://github.com/jrmazarura/wikipedia-sentence-selection.

Table 1: Top 25 topics from training set.

Topic Number	Coherence score	Topic words
780	-0.521	specie animal mammal group year early include large ago small
686	-0.565	millipede centipede specie order leg pair segment group body large
792	-0.581	bamboo culm flower shoot specie plant seed year grow time
162	-0.613	scorpion tarantula specie spider female leg male include large pressure
98	-0.661	poverty poor live income child people world population day social
164	-0.663	world theory concept term thing sense view understand form refer
628	-0.672	emotion theory experience individual emotional person social action behavior people
748	-0.675	band tour album rock record music member year group play
636	-0.694	dew form water surface small find night device include leave
155	-0.695	faisal saudi saud king country establish religious issue family royal
670	-0.713	leisure boy time activity sport family include male work increase
541	-0.726	duck tribe subfamily make family include breed water generally word
218	-0.727	cave form rock water include early find sea large deep
309	-0.736	giraffe neck long animal protein depend find high form percent
574	-0.741	chameleon catfish specie family tongue large sound signal include body
18	-0.744	kiln hut heat fire dry temperature wood build design type
771	-0.751	lice louse host body head bird specie egg live feed
551	-0.754	date year calendar start ad name reference early event begin
315	-0.755	boar specie wild butterfly area male year population large plant
704	-0.759	marshe water marsh plant pool form high occur habitat type
306	-0.770	perfection fly perfect number concept man art century body great
463	-0.771	material cut blade metal shape wood tool design type small
185	-0.782	pregnancy miscarriage week woman risk fetus birth increase age term
224	-0.785	health disease include medical condition care treatment factor study risk
334	-0.787	cattle zebu breed animal milk african originate carry region period

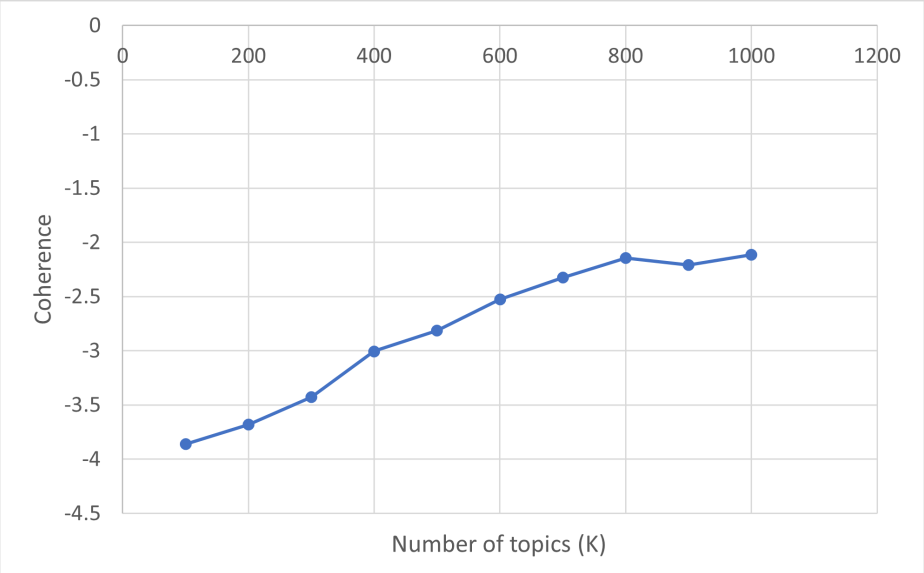

Figure 2: Average coherence score for $K = \{100, 200, ..., 1000\}$.

5.2 Document selection using trained topic model

The next step in the document selection process was to use the trained topic model to select articles from Wikipedia and then select sentences from them. For each article in Wikipedia, the model was used to determine the topic distribution and only articles satisfying the following conditions were retained:

1. The dominant topic of the article had to be amongst the 500 topics with the highest coherence scores from the trained model.
2. The dominant topic needed to make up at least a quarter of the article's topic distribution.

The 500 topic cutoff used in #1 was selected by assessing the coherence scores. Based on inspection, the interpretability of the topics appeared to noticeably deteriorate as the coherence scores decreased below -2. The first 500 topics had coherence scores above this value. In addition, the 500-topic cutoff also ensured that a broad variety of topics were covered. Choosing The 25% cutoff used in #2 not only ensured that the topic made up a significant proportion of the article's topic, but it was also necessary to ensure that only *some* of the articles from Wikipedia were selected.

After this step, the final collection of sentences was compiled by randomly selecting an equal number of articles from each topic and then randomly selecting one sentence from each article. For example, to create a collection of 5 000 sentences, 10 articles belonging to each of the 500 topics were randomly selected and one sentence was randomly selected from each of the articles. Only sentences with at least 15 words made up of 3 characters or more were selected. This was done to avoid retrieving sentences that were very short and did not contain much information.

6 Results

The final selection of sentences can be found at https://github.com/jrmazarura/wikipedia-sentence-selection. Table 2 shows a snippet of the sentences selected by the proposed method. From the table, it is evident that many topics do not appear to be related to the South African or even African context as desired. In selecting the articles with dominant topics belonging to our top 500 topics, it can be seen that many of the sentences are

about seemingly unrelated topics.

For example, consider Topic 218 in Table 1. This topic appears to be related to ocean caves. The trained topic model was used to retrieve related articles from Wikipedia and the titles of the articles were recorded (the list of titles is also available on GitHub). Some of the Wikipedia articles retrieved have titles such as *Cave*, *Stalactite* and *Stalagmite*. However, there are much more which seem less important in our context, such as *Velebit caves* (caves in Croatia), *Abrakurrie Cave* (caves in Australia) and *Ogof Agen Allwedd* (caves in Wales) amongst many others. The topic model was able to select 482 cave-related articles, but unfortunately, most of them are about specific caves from all over the world. Ultimately, this means that sampling random sentences from these randomly selected articles will most likely produce sentences that are unrelated to South Africa. The following are examples of some of the cave-related sentences that were selected:

"The Bull Thistle Cave Archaeological Site is an archaeological site on the National Register of Historic Places, located in Tazewell County, Virginia."

"The Caves of Hotton are speleothem caves located in Wallonia near Hotton in Belgium, which were discovered in 1958 and are around 5 or 6 km long and 70 metres deep."

This observation is likely due to the topic model trying to find topic distributions for over five million articles over only 800 topics. It is reasonable to assume that such a large corpus is likely to contain much more than 800 topics. In addition, topics such as the cave topic that was considered as an example, are likely to have many related subtopics, e.g. *African caves* or *ice caves*, etc. If there is only one topic related to caves, then all articles will be grouped into one topic creating a collection of many cave-related articles covering a broad scope.

7 Discussion & future work

Our first attempt at using topic modelling to identify Wikipedia articles that would be suitable for translation from English into the 10 other South African languages yielded an extremely diverse set of sentences, many of which contain words that are probably not lexicalised in any of the target languages. Training the topic models on the SIL-CAWL was clearly not adequate to avoid this challenge.

In the next phase of the research we will aim to find pruning criteria that could be used to "focus" the topic models on content that is more relevant to the South African context. We would also like to establish a relationship between the selected text and the resulting translation in the target language to ensure that both relevant and translatable text has been chosen. Such a measure will also allow a comparison between the proposed topic model approach to text selection and a fully random baseline.

Notes

[1] `https://repo.sadilar.org/`

[2] In this paper *text generation* only refers to text produced by humans. Automatic text generation will not be considered because the technique itself relies on the availability of text data, a resource that is not available in the circumstances relevant to this study.

[3] `https://radimrehurek.com/gensim/auto_examples/index.html`

[4] Perplexitiy measures a model's ability to predict new data. A low perplexity score is assumed to indicate good model performance.

Table 2: Selected sentences from Wikipedia

	Selected sentences
1.	The plateau originated in the Gondwanan breakup and is one of the five major submerged parts of Zealandia, a largely submerged continent.
2.	As such, it has become the site of the small Tabor Mountain Ski Resort, which is one of Prince George's two local ski hills, the other being the small Hart Highlands Alpine Park on the north side of the city.
3.	The stratovolcano lies above the regional Liquine-Ofqui Fault zone, and the ice-covered massif towers over the south portion of Pumalín Park.
4.	The sedimentary rock was more fragile than the metamorphic rock formed by the contact of the magma and the surrounding sedimentary rock.
5.	Examination by a severe weather team from the Bureau of Meteorology examined the damage in the Bucca and Kolan region and recorded it as an 'F4' on the Fujita scale.
6.	Formed in 1922 as the Westby Ski Club, the all-volunteer club held the first ski jumping tournament southeast of Westby, near Bloomingdale, Wisconsin in 1923.
7.	It is often considered to be one of the most spoiled of the Munros, due to the Glenshee Ski Centre which covers the eastern slope of the mountain.
8.	The de Lalande crater is named after the French astronomer Marie-Jeanne de Lalande (1768-1832), illegitimate daughter of astronomer Joseph Jerome de Lalande (1732-1807).
9.	The hill is 522 metres (1712 feet) high and is the highest point of the relatively low-lying county of Renfrewshire and indeed the entire Clyde Muirshiel Regional Park of which it is a part, having a considerable Topographic isolation.
10.	Additionally, a volcano was hypothesized to exist in the Nova Iguaçu area, in Rio de Janeiro, and was called the Nova Iguaçu Volcano.
11.	There is no precise definition of surrounding base, but Denali, Mount Kilimanjaro and Nanga Parbat are possible candidates for the tallest mountain on land by this measure.
12.	The Dease Plateau is a sub-plateau of the larger Yukon Plateau, and is located in far northern British Columbia, Canada, northwest from the Deadwood River to and beyond the Yukon-British Columbia boundary.
13.	The ruins on its top derive from the year 1449, when Oberhohenberg Castle together with the town of Hohenberg, at the foot of the mountain, were destroyed in a local feud.
14.	Enoggera Hill is a small mountain of the Taylor Range in Australia in the Brisbane suburb of Enoggera, in Queensland, with a peak of 273 meters (896 feet) above sea level.
15.	SnoCountry Mountain Reports was the first and is now the largest snow conditions reporting service in the world.

Acknowledgements

The localisation of the Mozilla Common Voice platform in South Africa is supported by the German Federal Ministry of Economic Cooperation and Development (BMZ), represented by the GIZ project FAIR Forward - Artificial Intelligence for All.

References

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Boyd-Graber, J., Hu, Y., Mimno, D. et al. (2017), 'Applications of topic models', *Foundations and Trends® in Information Retrieval* **11**(2-3), 143–296.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. & Blei, D. (2009), 'Reading tea leaves: How humans interpret topic models', *Advances in neural information processing systems* **22**.

Eiselen, R. & Puttkammer, M. (2014), Developing text resources for ten South African languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', European Language Resources Association (ELRA), Reykjavik, Iceland, pp. 3698–3703.
URL: *http://www.lrec-conf.org/proceedings/lrec2014/pdf/1151_Paper.pdf*

Goyal, N., Gao, C., Chaudhary, V., Chen, P.-J., Wenzek, G., Ju, D., Krishnan, S., Ranzato, M., Guzman, F. & Fan, A. (2022), 'The FLORES-101 evaluation benchmark for low-resource and multilingual machine translation', *Transactions of the Association for Computational Linguistics* **10**, 522–538.

Griesel, M. & Bosch, S. (2020), Navigating challenges of multilingual resource development for under-resourced languages: The case of the African wordnet project, *in* 'Proceedings of the first workshop on Resources for African Indigenous Languages', pp. 45–50.

Mimno, D., Wallach, H., Talley, E., Leenders, M. & McCallum, A. (2011), Optimizing semantic coherence in topic models, *in* 'Proceedings of the 2011 conference on empirical methods in natural language processing', pp. 262–272.

NLLB Team, Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., Jarrett, S., Sadagopan, K. R., Rowe, D., Spruit, S., Tran, C., Andrews, P., Ayan, N. F., Bhosale, S., Edunov, S., Fan, A., Gao, C., Goswami, V., Guzmán, F., Koehn, P., Mourachko, A., Ropers, C., Saleem, S., Schwenk, H. & Wang, J. (2022), 'No language left behind: Scaling human-centered machine translation'.

Pretorius, L. & Wolff, F. (2020), *Wikipedia as a Transformative Multilingual Knowledge Resource*, Cambridge University Press, p. 232–246.

Rana, T. A., Cheah, Y.-N. & Letchmunan, S. (2016), 'Topic modeling in sentiment analysis: A systematic review.', *Journal of ICT Research & Applications* **10**(1).

Snider, K. & Roberts, J. (2004), 'SIL comparative African wordlist (SILCAWL)', *Journal of West African Languages* **31**(2), 73–122.

Voicing in Ngamambo: A Descriptive perspective

Nyindem, Sirih-Nagang Nancy
University of Bamenda, North West Region

Asongwed Tah (Retired Lecturer)

Abstract

This paper describes voicing in Ngamambo, a semi Grassfields Bantu language in the North West Region of Cameroon. The language is classified under the Momo sub-language family (Eberhard, David M., Gray F. Simons and Charles D. Fenning, 2020). Ngamambo is unwritten, and research on the language is scanty. The only available literature on the language is by Asongwed & Hyman (1976)), Achiri-Taboh (2014) and Lem Atanga (2020) However, there has been some recent attempt by the Mbu Language Committee (MLC) to study the language. Interest in the study of Ngamambo stems from the imperative of undertaking a comprehensive description of the language. Preliminary research has revealed the existence of voicing in the language. Voicing is a process whereby the pronunciation of a word is influenced by one of the sounds. Data was obtained from Ngamambo native speakers (informants) over six months. The originality of this study resides in the fact that very little research has been carried out on the language. The authors of this paper discuss one aspect of the language and hope that subsequent studies will determine if voicing is also present in other Grassfields languages, especially the Momo sub-language family. The phonological process of voicing in Ngamambo has been observed when a voiceless sound becomes voiced depending on the environment. It is hoped that understanding this phenomenon would lead to a better understanding of voicing related to language learning.

Keywords: Voicing, Ngamambo, Standardisation

1 General Introduction

The study of voicing is essential in that it helps in the better understanding of connected speech processes. One of the primary goals of linguists is to describe intriguing phenomena in human languages. This paper describes voicing in Ngamambo as the process that takes place when two consonants with different phonological specifications for voicing occur in a word structure and influence the quality (sonority) of the vowels that precede them.

This is a descriptive study. The authors of this paper are members of the Mbu Language Committee (MLC) which was established in early 2021 to study Ngamambo. The MLC comprises 15 members (6 females and 9 males) ranging in age from 35 to 75 years. They are fluent native speakers of the language and provide the data which the two linguists use to analyse the language.

The Committee meets twice a month. Focused group discussions take place regularly to verify and cross-check the data collected.

The MLC has so far presented a proposed Ngamambo alphabet and a writing system for vetting to the Ngamambo-speaking community through two-well attended Mbu worldwide community zoom meetings.

This paper is divided into two sections: section 1 is a general introduction to the phenomenon of voicing, and section 2 discusses the locus of voicing in Ngamambo.

Research has shown that voicing usually occurs during fast speech (Nollan & Holst, 1993). In Ngamambo, the change in the voice quality of sounds occurs in both fast and slow speech. The target of the investigation is voiced and voiceless intervocalic stops at the word-final and phrase level. The research data is based on a recording of 13 native speakers of Ngamambo. The data is analysed in three blocks focusing on (1) word-final stops, (2) voicing at the word boundary and (3) voiceless process across the word boundary.

1.1 The concept of voicing

In phonology, voicing (or sonorisation) is a sound change where a voiceless consonant becomes voiced due to the influence of its phonological environment (Consonant voicing and devoicing, 2022). Voicing occurs when the larynx's adductor muscles close the larynx, enabling the inner edges of the vocal cords to be in light contact. Pressure from the pulmonic

airstream builds up, and subglottal pressure rises. As an effect, the vocal cords are blown slightly apart, and the compressed air below the glottis then flows through the narrow gaps in the larynx at a very high speed due to high subglottal pressure.

Voicing in sonorant consonants involves the escape of air freely through the oral and the nasal cavities, whereas voicing in obstruent involves the passage of air from the oral cavity partially or fully blocked, leading to a rapid build-up of supraglottal air pressure (Grijzenhout 2000, Dmitrieva 2014).

Stop consonants are produced by a complete closure in the vocal tract. The complete oral closure in oral stop is combined with a velic closure that prevents the air from escaping through the nasal cavity. As such, there is a rise in intra-oral pressure so that, when the closure is released, the compressed air escapes to the atmosphere with a stop burst, and pressure falls rapidly. According to Warren (1996), voiceless oral plosives have a volume of air of approximately 50ml. Also, voiced stops are always shorter than voiceless stops, and if voiced stops were as long as voiceless ones, then voicing would die out, since the subglottal and oral pressure would equalise. In summary, for voicing to take place, the vocal folds have to be in contact, and there has to be a sufficient pressure drop across the glottis.

1.2 Voicing in Ngamambo

Segments are not articulated in isolation in speech; rather, segments influence each other and are sensitive to context. Voicing is a phonological process that explains the varying qualities in sound production. It is the modification of the phonological features of a segment due to the influence of an adjacent segment. Different commands are given to the glottis to produce segments with different phonological specifications for voicing in running speech. In Ngamambo, we observe that a voiceless consonant becomes voiced at the word boundary. This process seems to be part of the native speaker´s intuition of the grammar of the language. For instance, in the "infinitive" form of the verb, the final segment is voiceless, and when, for example it is conjugated to mark tense, the segment becomes voiced at the word boundary. The examples below illustrate this phenomenon.

Table 1: Verbs ending in /k/ and with /ɔ/ preceding the /k/

Verb (root)	**Gloss (infinitve form)**	**Conjugated verb form**	**Gloss (progressing tense)**
gɔ̀k	"to fall"	gùgə́	"falling"
fɔ́k	"to clean"	fúgə́	"cleaning"
tɔ̀k	"to spit out."	tùgə̀	"spitting out"
zɔ́k	"to hear"	zúgə́	"hearing"

Voicing occurs in Ngamambo when there is a boundary between the segments. When a voiceless consonant at the word boundary is followed by another segment, the voiceless sound becomes voiced. For instance, in the word gɔ̀k # gùgə́, the final plosive of the root form /k/ changes to /g/. As observed from the data above, this voicing phenomenon is expressed in this context when the verb form is conjugated to mark the progressive tense. Also, the vowel of the root verb changes from a mid-low back-rounded vowel /ɔ/ to a high-back round vowel /u/.

Table 2. Nouns ending in /k/ and with /ɔ/ preceding the /k/

Nouns	**Gloss**	**Noun +Mod**	**Gloss**
útsɔ́k	"mouth"	útsúg mə́t	"his/her mouth"
ndzɔ́k	"kind of peanut"	ndzúg zé	"that kind of peanut"
bɔ́k	"dog"	búg zé	"the dog"
dzɔ́k	"honey"	dzúg zé	"the honey"
ə́tɔ́k	"head"	ə́túg zé	"his/her head"

Table 2 above shows that nouns ending in /k/ and preceded by /ɔ/ follow the same rule as verbs inflected for progressive tense. The

common denominator in both cases is that the consonant change takes occurs across a word boundary and affects the preceding vowel.

Table 3: Verbs ending in /k/ and with / ə̀/ preceding the /k/.

dzə́k	"to eat"	"dzɨgə́"	"eating"
sə́k	"to slice"	sɨgə́	"slicing"

In the case of verbs ending in /k/ and preceded by /ə/, when conjugated in the present progressive tense, the /k/ changes to /g/ and the middle low vowel /ə/ changes to a back /ɨ /.

Table 4: Nouns ending in /k/ and preceded by /ə́/.

Noun	**Gloss**	**Noun +Mod**	**Gloss**
ɨbə́k	"camwood"	ɨbɨk wé	"the camwood"
lɨdzə́k	"food"	lɨdzɨk té	"the food"

Where the noun ends in /k/ preceded by /ə́/ across a word boundary, the /k/ changes to /g/ and the /ə́/ changes to /ɨ/.

Table 5: Words (nouns and verbs) ending in /y/. Consider the following:

sɛ̀y	"to be selfish"	ú wé sìgɛ̀	"he is selfish"
sɛ́y	"ground"	síg zé kà	"the ground is hard"
dɛ̀y	"to cry"	ú wé dìgɛ̀	"he is crying"
fúbɛ́y	"knife"	fúbíg fɛ́	"the knife"
ə́tʃwɛ́y	"sun	ə́tʃwíg zɛ́	"the sun"
kyɛ̀y	"song	kyíg wɛ́	"the song"

Table 5 shows that voicing in Ngamambo is not limited to /k/ in word-final position changing to /g/ across word boundaries and causing a change in the vowel quality from /ɔ/ to /u/.

The examples show that /y/ in word-final position changes to /g/ and the preceding vowel /ɛ/, a front, mid low vowel changes to /i/, a front high vowel as in the case of /k/ where the preceding back mid low /ɔ/ vowel changes to the back high /u/. The data so far shows that the only vowel that precedes /y/ is /ɛ/ which, as shown, changes to /i/ while /y/ at the same time changes to /g/.

Table 6: Verbs ending in /t/

Verb (root)	**Gloss (infinitive form)**	**Conjugated verb form**	**Gloss (progressing tense)**
wàt	"to cut"	mə́ wé wàrə̀	"I am cutting."
gàt	"to leave"	mə́ wé gàrə̀	"I am leaving"
kə̀t	"to hang (dress)	mə́ wé kə̀rə̀	hanging
kót	"to tie"	mə́ wé kórə́	"tying"
phút	"to eat"	mə́ wé phúrə́	"eating"
nyìt	"heavy	mə́ wé nyìrə̀	
tʃwét	"to survive"	tʃwéré	"surviving"

From the data in Table 6 it is noted that in Ngamambo, when verbs ending with /t/ are conjugated to mark the progressive tense, the voiceless alveolar oral stop /t/ changes to a fricative /r/. However, there is no concomitant change in the vowel quality as in the case of words ending /k/ In the case of /t/, the vowel in the root form of the verbs does not change.

Table 7: Nouns ending in /t/

ɲát	"deer"	ɲát zé	"the deer"
kə́t	"penis"	kə́t wé	"the penis"
ɲót	"body"	ɲót wé	"the body"
kát	"wheel"	kát wé	"the wheel"
ə́tə́t	"mad man"	ə́tə́t zé	"the mad man"
tìt	"louse"	tìt zé	"the house"

The above examples in Table 7 show that nouns ending in /t/ do not change across word boundary, and there is also no vowel change. It was shown in table 5 that verbs ending in /t/ change when they are conjugated in the present

progressive tense, but the vowel remains the same.

Table 8: Words ending in /p/

Verb (root)	Gloss (infinitive form)	Conjugated verb form	Gloss (present progressive tense)
bɔ́p	"to be rotten."	ú wé bɔ́bé	"he is rottening"
tɔ́p	"to stir"	ú wé tɔ́bé	"he is stirring."
sóp	"to cut"	ú wé sóbé	"he is cutting"
zɔ̀p	"mourn"	ú wé zɔ́bé	"he is mourning"
káp	"to pluck"	ú wé kábə́	"he is plucking"

The data in Table 8 shows words ending with the voiceless bilabial plosive /p/. This voiceless plosive /p/ changes to its voiced counterpart /b/ when the word is conjugated to mark tense. It behaves the same way as verbs ending in /t/.

Table 9 above further confirms the existence of voicing in Ngamambo in that the final voiceless consonant /p/ in word final position changes to its voiced counterpart /b/ across a word boundary irrespective of the preceding vowel.

1.3 The locus of the trigger of voicing

This paper has adduced evidence to show that words ending in /t/,/p/ and /k/ undergo some form of voicing. However, there seem to be two types of voicing: one involving /t/ and /p/ where, when they are in final position, they maintain their voiceless character. However, across word boundaries, they become voiced but without a change in the vowel (quality or sonority of the vowel). It is proposed that this form of voicing in Ngamamba be called partial voicing since it only involves the consonant change.

The other form of voicing involves /k/ and /y/ in word-final position and their change to /g/ across word boundaries. In this type of voicing, the /ɛ/ infront of /y/ changes to /i/, the /ɔ/ in front of /k/ changes to /u/ and the /ə/ in front of /k/ changes to /ɨ / across a word boundary.

It is proposed that this second type of voicing be called full voicing because it entails both a consonant and vowel change.

Table 9: Nouns ending in /p/

Root (noun)	Gloss	Noun + demonstrative	Gloss
ŋgə̀p	"a type of calabash"	ŋgə̀b zé	"that type of calabash"
ətsɔ́p	"a curse"	ətsɔ́b zé	"that curse"
ŋgúp	"chicken/fowl"	ŋgúb zé	"that chicken"
káp	"money"	ǂkáb wé	"that money"
əzóp	"name of a hill"	əzób zé	"the Ezop hill"
ə́kóp	"raffin palm fr	ə̀kób zé	"the type of kolanut"
mbəp	"huckleberry"	mbə̀b zé	"the huckleberry"

1.4 Locus of the Trigger

The question arises as to whether the consonants /k/ and /y/ trigger the vowel change across word boundaries or the vowel triggers the consonant change across word boundaries. The locus of the trigger of voicing in Ngamambo would seem in some cases to be dependent on the class of the word; whether it is a verb or a noun as has been shown in various examples.

Furthermore, it would seem from the data that in some cases the vowel triggers the consonant change rather than the other way round where one might also posit that the consonant triggers the vowel change. For example,

Table 10

kə̀t	"to hang (dress)
kə́t	"penis"
káp	"money"

The argument in this case is that the word-final consonant generally remains constant until the vowel that precedes it changes. Usually the vowel that precedes the word-final consonant is not high. However, when it changes to become high, the word-final consonant (generally a voiceless consonant) changes to its voiced counterpart to "harmonize" the sonority of the new high vowel that has taken the place of the lower vowel.

Word category

From the data in this paper, it is clear that voicing occurs with verbs when inflected to mark tense (progressive tense). It also occurs with nouns. Further research will be undertaken to better understand the locus of the trigger of voicing in Ngamambo.

Devoicing

While this paper is on voicing in Ngamambo, it is important to point out that there are instances of devoicing in the language. It has been shown that some generally voiceless consonants in word-final position become voiced across word boundaries and the vowel quality of the vowels that were in front of them also change. Devoicing occurs where sounds lose their voicing quality when they are either preceded or followed by specific consonantal or vowel segments (Fujimoto, 2015). This seems to be the case in Ngamambo. Consider the example below:

Table 11: Verbs with /ə/

fétí	"to gather"	fété	"gathering"
sètì	"to tear"	sètè	"tearing"
jì'tì	"to grumble"	jì'tè	"grumbling"
nyí'tí	"to poison"	nyí'té	"poisoning"
bèrì	"to own"	bèrè	"owning"

From the data in Table 11, the quality of the vowel (the sonority of the vowel) changes from a front high unrounded vowel /i/ to a front mid-high unrounded vowel /é/ which shows that the high vowel /í/ has lost some of it high quality to become a mid vowel /é/.

1.5 Conclusion

This paper has attempted to describe the phonological phenomenon of voicing in Ngamambo. It has analysed the manifestation and trigger of the locus of voicing in the language. The study establishes that word-final voiceless consonants become voiced at a word boundary. Furthermore, the effect of the voicing is manifested in the quality of the vowels that precede the consonants.

This research is work in progress and it is hoped that further research will shed more light on the phenomenon of voicing in Ngamambo.

References

Adda-Decker, M. & Lori L. (1999). Pronunciation variants across system configuration, language and speaking style. Speech Communication 29 (2-4): 83-98.

Best, C. T. & Pierre Hallé. (2010). Perception of initial obstruent voicing is influenced by the gestural organization. J. Phonetics 38: 109-126.

Consonant voicing and devoicing. (2022, October 9). In Wikipedia. https://en.wikipedia.org/wiki/Consonant_voicing_and_devoicing.

Coretta, S. (2019). An exploratory study of voicing-related differences in vowel duration as a compensatory temporal adjustment in Italian and Polish. Glossa: A Journal of General Linguistics, 4(1), 125. DOI: http://doi.org/10.5334/gjgl.869.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. Phonetica 22, 129–159. doi: 10.1159/000259312.

Dawood, H. S. A. & Ahmad A. (2015). Assimilation of Consonants in English and Assimilation of the Definite Article in Arabic. American Research Journal of English and Literature, Volume 1, Issue 4. ISSN 2378-9026.

Keating, P. A. (1984). Phonetic and phonological representation of stop voicing. Language. 60:286–319.

Dmitrieva, O. (2014). "Final voicing and devoicing in American English". The Journal of the Acoustical Society of America. 136 (4): 2174. Bibcode:2014ASAJ..136.2174D. doi:10.1121/1.4899867.

Fujimoto, M. (2015). 4 vowel devoicing. In H. Kubozono (Ed.), Handbook of Japanese phonetics and Phonology (PP. 167.214). Berlin, Müchen, Boston: De Gruyter Mouton. https//doi. Org/10.1515/9781614511984.167.

Grijzenhout, J. (2000). "Voicing and devoicing in English, German, and Dutch: Evidence for domain-specific identity constraints". CiteSeerX 10.1.1.141.5510

Kulikov, V. (2012). "Voicing and voice assimilation in Russian stops." PhD (Doctor of Philosophy) thesis, University of Iowa. https://doi.org/10.17077/etd.r6ib0d07.

Hallé, P. & Martine A.-D. (2012). Voice assimilation in French obstruents: A gradient or a categorical process? Tones and features: A festschrift for Nick Clements, De Gruyter, pp.149-175, 2011. ffhalshs-00684437.

Wetzels, W. L. & Joan M. (2001). The typology of voicing and devoicing. Language. 77:207–244.

Unpacking the Possibilities of a Vernacular Language Archive

Fagan, Henry
Fagan.henry@gmail.com
McNulty, Grant
grant.mcnulty@uct.ac.za
Hamilton, Carolyn
carolyn.hamilton@uct.ac.za
Suleman, Hussein
hussein@cs.uct.ac.za
Archive and Public Culture Research Initiative

Abstract

The Five Hundred Year Archive is a research project of the Archive and Public Culture Research Initiative based at the University of Cape Town. In an effort to stimulate greater engagement with the deep southern African past, the project has created a corpus of vernacular resources ranging from the earliest available to 1910. It includes productions by an array of African intellectuals in a host of African languages. The vernacular corpus, with its rich metadata, constitutes an extended language and conceptual archive. It is useful to historians, but may also offer research possibilities in other fields, particularly if used in conjunction with contemporary computational methods.

Keywords: African Intellectuals, Vernacular Writing, Orthography, Metadata, Copyright

1. Introduction

The Five Hundred Year Archive (FHYA) [1] is a project of the Archive and Public Culture Research Initiative (APC) [2] at the University of Cape Town (UCT). It seeks to stimulate engagement with the deep past - the neglected eras of the southern African past before the advent of European colonialism. One of the main challenges of conducting research in this area is the severe lack of material, including written sources. What written sources there are have complex histories of production that need to be researched in their own right. While researchers might make use of non-textual material like objects and sound recordings, much of this is misidentified, undated, lost, disorganised or scattered in institutions across the world, making the material difficult to access and use. Working across media and disciplines, with local and international partner institutions, we at the FHYA have created a handful of digital projects that address these challenges.

2. Convening an archive for the deep past

One of these projects is the recently launched online platform, EMANDULO [3], which digitally convenes a growing collection of historical materials relevant to the deep southern African past and makes them available through a single, searchable database. The FHYA selects materials from existing archival or museum collections and galleries, as well as from personal collections. Partner institutions include the Wits University Historical Papers, the Johannesburg Art Gallery, the KwaZulu-Natal Museum, the Amafa/Heritage KwaZulu Natal provincial heritage conservation agency, the Swaziland National Archives, the Killie Campbell Africana Library, the Cambridge Museum of Archaeology and Anthropology, the Austrian Academy of Sciences, the Bews Herbarium at the University of KwaZulu-Natal and the Cambridge University Library. Sometimes individuals or institutions offer digitised or analogue materials to the FHYA. The materials range from hundreds of published texts to recorded oral histories, excavated objects, reports, theses, maps, photographs, images, and researchers' notes. The curation of these materials online offers an opportunity to reorganise and re-position the materials in ways that are designed to challenge the ways in which they have been framed historically by colonial knowledge practices.

The holdings include early productions of the past by African intellectuals who were brokering the history of past generations into the new colonial world. Many are texts written in vernacular languages, which scholars have generally ignored

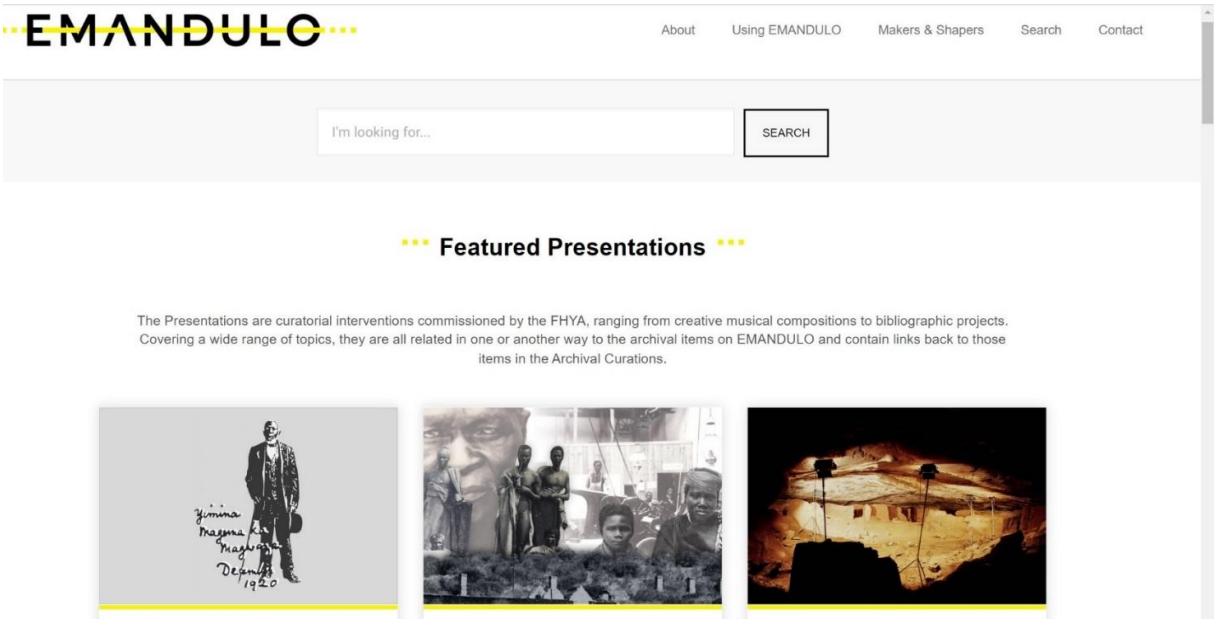

Figure 1: A screenshot of EMANDULO's home page.

as historical sources. On EMANDULO these materials are either presented in themed foci (see for example EMANDULO's collection of materials related to Magema Magwaza Fuze) [4] or are added to the "General Repository" (GR), which includes materials added both by the FHYA team (FHYA Depot) [5] and by registered contributors (Public Depot) [6].

Items in both the GR and the themed sections, like the Fuze archive, have extensive, searchable metadata which tracks, as far as possible, their archival histories - provenance and collection histories - foregrounding the changes to which they have been subjected over time. Provenance deals with the creator or originator of materials before they were lodged in an institutional setting while collection histories refer to the multiple hands that shaped them in each institutional context.

In this paper we focus on the vernacular texts on EMANDULO in both the themed sections and in the GR. We draw attention to the remarkable opportunities that they open up, not only for historians and historical linguists, but as vernacular resources for teachers and the many as yet unknown ways in which they might be used. Indeed, our aim in bringing a paper to this workshop is to use it as a prompt for wider discussion with a spectrum of experts about further areas of possible development and use.

3. Building the Vernacular Corpus

3.1 Selecting Materials

The FHYA selects materials that are pertinent to the five-hundred-year period before European colonisation in what is today the Eastern Cape, KwaZulu-Natal, parts of the Highveld, eSwatini and Lesotho, deliberately extending across modern-day borders that did not exist in the eras before colonialism. The vernacular texts, rendered in a variety of orthographies, occur in early versions of what later became standardised isiZulu, isiXhosa, seSotho, and siSwati. The corpus also contains texts in dialects such as 'siNguni' (Ndwandwe), sePhuthi and isiBhaca. It includes, amongst others, works by well-known writers like John Dube, R.R.R. Dhlomo, Magema Fuze, S.E.K. Mqhayi, Henry Masila Ndawo, and Thomas Mofolo, and many lesser known writers. The texts touch on a broad range of subjects ranging from histories, biographical writing and

historical fiction, to plays, social commentaries, and theology. Importantly, the corpus includes early dictionaries.

We aim to include on the site *all* the vernacular texts we are able to locate from the earliest times to 1910. For the period after 1910 we include texts that for any number of reasons we, or our registered contributors, consider relevant to the study of the deep past in southern Africa.

3.2 Locating Vernacular Materials

The FHYA proactively seeks out texts by combing catalogues and databases (such as UCT Libraries or WorldCat) for keywords between particular dates. This enables us to locate appropriate materials, particularly published books. It also scours open-access repositories such as Google Books or the Internet Archive for texts whose copyright has expired. Suitable online finds are subsequently downloaded and added to EMANDULO. In many instances, the texts are contributed, or pointed to, by scholars who locate them in the course of their research. Before the UCT fire (Davids, 2021) we were able to get physical access to early books in the UCT African Studies Collection, which we then digitised. We sometimes access relevant books in other collections.

3.3 Digitising Materials

Where necessary the FHYA digitises the materials it acquires and creates PDFs. Most scans are done at a moderate quality and in black and white to ensure their file sizes are not prohibitively large when uploading them to EMANDULO. In rare cases, such as with books where visuals are the focus, the scans are made in colour.

The FHYA applies optical character recognition (OCR) to all texts, making them machine readable and thus searchable on EMANDULO. Although the FHYA's scanner is capable of recognising characters in several languages, vernacular languages (aside from isiXhosa) are not intrinsically recognised by the software. There are also certain things that undermine the accuracy of the OCR. These include materials where the text is faded, those with rare and elaborate old fonts, and outdated or non-standard orthography. OCR problems double as search problems, as EMANDULO's search function cannot correctly index and search the affected text. For items where the text is difficult to discern, the FHYA has begun to introduce typescripts, where full text is made accessible as a related document. In some instances, handwritten texts have typed up summaries produced for one or another purpose in the past. Sometimes transcripts or summaries are contributed by registered users. Consequently, the original items are indirectly searchable.

Some items are old and fragile, which makes scanning them difficult. This may necessitate imprecise scanning (such as slightly skew pages) specifically undertaken to prevent straining the pages or the frames of delicate books. Many materials are already damaged and might have torn or missing pages, or faded text. In such cases, where possible, the FHYA will incorporate the missing page from a better quality copy or a later edition of the production. Such interventions are, of course, carefully noted in the metadata.

3.4 Adding Metadata

Materials sourced by the FHYA for the GR include all known details about each particular item. The FHYA team then researches and adds additional information to the metadata of a particular item so that users of EMANDULO know its archival history – when it was created, by whom and under which conditions, and what happened to it over time. In each case the source of the metadata is provided. Enriched metadata records thus make visible the various 'hands' involved in making and shaping the material over time. This also allows for different copies of the same item (with unique archival histories) to be acknowledged individually. In other words, it is the FHYA's emphasis on metadata that allows it to treat its materials as archival objects.

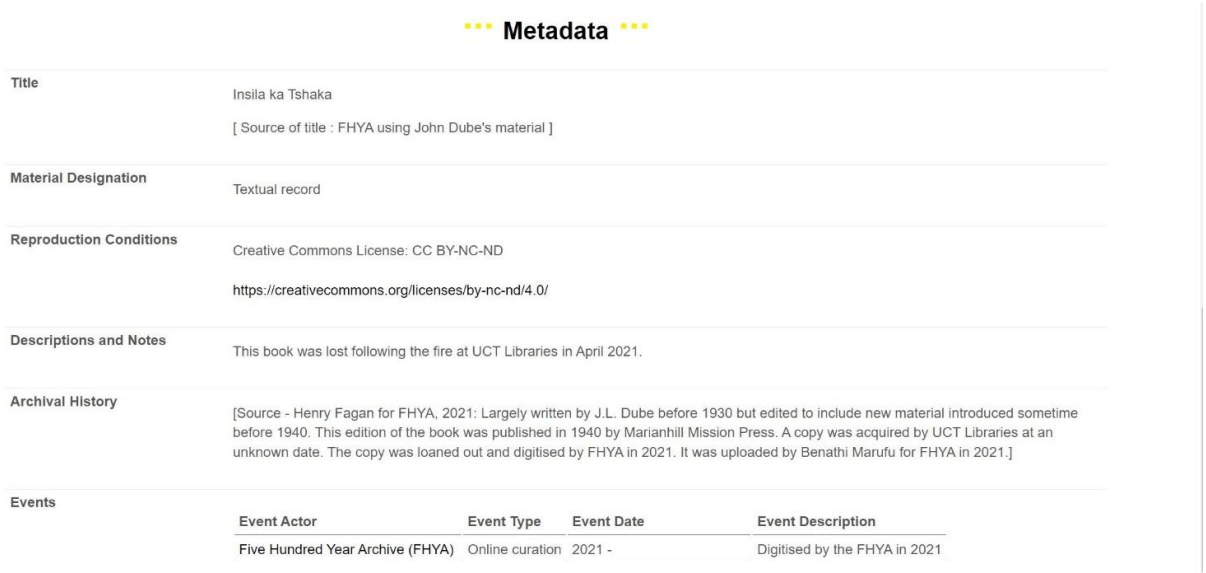

Figure 2: A screenshot of the metadata of an individual item.

The FHYA records metadata meticulously and in a consistent style based on the International Council on Archives' General International Standard Archival Description (ICA, 2000). The ICA-Atom metadata format has an emphasis on encoding archival processes related to objects and collections. The Europeana Data Model (EDM), which also focuses on cultural archives, extends that notion to cater for the principles of Linked Open Data, with the intention of supporting distributed interoperable archives. The specific metadata format used in EMANDULO is, however, not crucial, as the key elements can easily be mapped between ICA-Atom and EDM.

At the systems level, Europeana and EMANDULO are driven by very different objectives, with the former designed for costly large scale distributed collection management and the latter for resilient smaller collections in low-resource environments.

It is sometimes necessary to enter metadata that departs from the characteristics of the physical object. For example, colonial-era books with titles that are extremely offensive in a contemporary context are slightly altered (by using inoffensive substitutes) such that their titles are not explicit but still recognisable. Problems with metadata accuracy can arise when items are mislabelled by institutions, or when additional data about the materials cannot be located. In such cases, the FHYA looks to clarify and put down as much useful information as it can. The final step entails uploading material to EMANDULO, where it is accessible to the public. EMANDULO users can browse the items' metadata and search through metadata and OCR-generated texts.

3.5 Copyright

For the General Repository, the FHYA focuses on published texts that are no longer subject to copyright. Copyright for archival materials, like those housed in the Fuze archive, is less clear cut. We have received a legal opinion that if an archival document has been available in a public institution for 50 years or more, it can be treated in the same way as published material. In certain instances, it is necessary for the FHYA to apply to holders of copyright for permission to upload material. Occasionally, institutions may claim copyright over materials to which they may not have a legal right. In cases where the copyright of the material is unknown but the item is conjectured to be in the public domain, the FHYA uploads the item with a short disclaimer outlining that it will be taken down should there be a valid copyright assertion. All materials uploaded to

EMANDULO are available under a Creative Commons license (CC BY-NC-ND 4.0).

Another issue arises when institutions produce and then exhibit digital copies of materials but refuse to share those digital copies with the public on the basis that they are a production of the institution. In circumstances such as these, public institutions exert controls that may not be in the public interest. The FHYA works to generate debate and discussion when such cases arise, with a view to shifting policies.

4. A vernacular language and conceptual archive

The FHYA has made a point of locating and uploading to EMANDULO sources in local African languages and allows registered users to do the same. We value contributions from a broad spectrum of users, from professional academics to those with an interest in history and culture. Registered users acknowledge that they either own the copyright of contributed material or that it is out of copyright and that there is nothing racist, sexist or homophobic in what they contribute. All contributions are moderated before being made publicly available. The collected texts form a rich and specifically tailored corpus - one that will grow as more resources are added. Given the FHYA's attention to the production of rich metadata, as much as possible that is known about the material - its dates of production, of publication, etc. - is easily established. This makes it possible to track how vernacular languages were being used at particular points in time and how language use has evolved.

A new generation of scholars working in the APC are paying productive attention to concepts within vernacular writing, and more specifically, concepts in motion across time. As these concepts are infused with particular connotations and meanings, they are not easily translated, but attention to their vernacular usage can bring their historical 'lives' into view. The term 'umbuso' in what is now the KwaZulu-Natal region offers a case in point. The nineteenth century saw dramatic changes in the nature of political power in the region as independent rulers were replaced by colonial governments and the chiefs they appointed. The earliest African writers, using local languages, recorded their understandings of what umbuso was in the eras immediately before their own time. In part this was because they were concerned to engage critically with newly introduced colonial ideas about rule, which the colonial authorities represented as being based on traditional African rule.

Terms like umbuso that were previously used to refer to significantly different forms of rule, were pressed into service in communication about new forms of colonial rule. While the term stayed constant over time, its meaning was changing. The searchability of carefully dated texts in our corpus makes it possible for historians to track the changing meanings of the term over time. What is considered by many today to be a traditional form of rule by chiefs can thus be shown to have a specific, changing history across time. As EMANDULO helps to illustrate within the southern African context, shifts in meaning and orthography take place as political contexts evolve.

Standardised forms of African languages often coalesced around mission stations in particular geographical centres. For example, the Lovedale Missionary Institute situated in what is now the Eastern Cape region became a hub for the production of isiXhosa writing during the early twentieth century (Peires, 1980). Likewise, *Ilanga Lase Natal*, a newspaper founded by John Langalibalele Dube and Nokutela Dube, became a spring-board for early isiZulu writing and isiZulu intellectual discourse (Hughes, 2011). By bringing together materials that show dialectical and orthographic variations, the FHYA corpus lends itself to ongoing work in mapping and periodising these developments. Subject to careful historicisation and contextualisation, the kinds of resources found in the GR are green fields for those interested in the past and in historic changes in orthography and the meaning of terms and concepts, especially when used in conjunction with contemporary computational methods. In

this respect the FHYA corpus of digitised vernacular texts constitutes an extended language and conceptual archive.

5. Early Computational Experiments using FHYA data

The corpus is already proving useful to computer scientists seeking to develop machine-based interventions to advance new research possibilities. As the initial archive was developed, a number of different research problems were identified and explored at the intersection of the digital humanities and computer science.
The text and image collections were used as the basis for experiments with search technology to determine how end users react to multimodal search - where users can enter textual queries or submit pictures, or both - and if specific common search algorithms are applicable to historical South African texts (Singh, 2022). It was determined that some techniques (like stemming) are applicable, but others (like thesaurus use) require specialised development. In addition, users showed a preference for textual search where concepts were abstract but image search where concrete visual representations were possible.

As metadata was being created by the FHYA team, it also became apparent that linking entities (e.g. people, places) where there is orthographic variation is a human-intensive task. Some computational linguists have recently explored the use of automated Named Entity Linking based on machine learning techniques. Experiments were therefore conducted to test the applicability of machine learning techniques to disambiguate vernacular representations of names in FHYA documents and metadata using statistical language models such as BERT and XLM-R (Dunn, 2022). Results from this work have highlighted that techniques that work well for European languages need further refinement to adequately handle the morphological characteristics of Nguni languages.

These early experimental studies demonstrate how the corpus developed by the FHYA is an enabler of research in computational disciplines and establishes a necessary symbiotic relationship between collection development and algorithm development in vernacular languages.

6. Conclusion

Our vernacular language corpus - an extended language and conceptual archive - recognises variability, fluidity, and historical linkages between non-standard forms of language across time. It is proving to be a useful teaching resource and we expect it to facilitate historical research. As the possibilities of our vernacular corpus extend beyond the confines of the discipline of history, we seek insights from other disciplines concerning its potential.

We are especially interested in finding out to what extent, and how, this corpus might be of interest to linguists and language specialists. What might we do to improve its usability for researchers and teachers outside of history? To what extent is the capacity of the corpus to track orthographic changes and the emergence of standardisation of interest, and can we enhance the ability of the system to facilitate this in any way?

Notes

[1] See http://www.apc.uct.ac.za/apc/research/projects/five-hundred-year-archive

[2] See http://www.apc.uct.ac.za/

[3] See http://emandulo.apc.uct.ac.za/

[4] See http://emandulo.apc.uct.ac.za/metadata/Fuze/index.html.

[5] See http://emandulo.apc.uct.ac.za/metadata/FHYA%20Depot/index.html

[6] See http://emandulo.apc.uct.ac.za/metadata/Public%20Depot/index.html

Acknowledgements

This research was made possible with the support of the FHYA team and the APC's broader network of researchers and students. Special thanks to Benathi Marufu for her work with the vernacular materials.

References

Archive and Public Culture Research Initiative 2022, EMANDULO, viewed 22 August 2022, http://emandulo.apc.uct.ac.za/

Davids, N 2021, 'Reflecting on the devastating UCT fire', *University of Cape Town News*, viewed 23 August 2022, https://www.news.uct.ac.za/article/-2021-06-23-reflecting-on-the-devastating-uct-fire

Dunn, JWD 2022, *Evaluating Automated and Hybrid Neural Disambiguation for African Historical Named Entities*, Master's dissertation, University of Cape Town, Cape Town.

Hamilton, C and McNulty G 2022, 'Refiguring the Archive for Eras before Writing: Digital Interventions, Affordances and Research Futures', *History in Africa*, first view, pp. 1-27.

Hughes, H 2011, *First President: A Life of John L. Dube, Founding President of the ANC*, Jacana Media (pty) ltd, Johannesburg.

International Council on Archive 2022, *ISAD(G): General International Standard Archival Description*, viewed 1 September 2022, https://www.ica.org/en/isadg-general-international-standard-archival-description-second-edition

Morelli, E 2022, *Exploring uMgungundlovu*, viewed 23 August 2022, https://studio-emandulo.uct.ac.za/fhya-exploring-umgungundlovu/

Peires, J 1980, 'Lovedale Press: Literature for the Bantu Revisited', *English in Africa* 7, no 1, pp. 71-85.

Ramji, H 2022, *Nongqawuse and the Great Xhosa Cattle Killing* , viewed 23 August 2022, https://studio-emandulo.uct.ac.za/nongqawuse-and-the-great-xhosa-cattle-killing/

Singh, SH 2022, *Investigating user experience and bias mitigation of the multi-modal retrieval of historical data*, Master's dissertation, University of Cape Town, Cape Town.

Izindaba-Tindzaba: Machine learning news categorisation for Long and Short Text for isiZulu and Siswati

Madodonga, Andani
Department of Computer Science, University of Pretoria, South Africa
andanim412@gmail.com

Marivate, Vukosi
Department of Computer Science, University of Pretoria, South Africa
vukosi.marivate@cs.up.ac.za

Adendorff, Matthew
Open Cities Lab
matthew@opencitieslab.org

Abstract

Local/Native South African languages are classified as low-resource languages. As such, it is essential to build the resources for these languages so that they can benefit from advances in the field of natural language processing. In this work, the focus was to create annotated news datasets for the isiZulu and Siswati native languages based on news topic classification tasks and present the findings from these baseline classification models. Due to the shortage of data for these native South African languages, the datasets that were created were augmented and oversampled to increase data size and overcome class classification imbalance. In total, four different classification models were used namely Logistic regression, Naive bayes, XGBoost and LSTM. These models were trained on three different word embeddings namely Bag-Of-Words, TFIDF and Word2vec. The results of this study showed that XGBoost, Logistic Regression and LSTM, trained from Word2vec performed better than the other combinations.

Keywords: South African native Languages, Low Resources Languages, Data Augmentation, Topic Classification, News Categorisation

1 Introduction

Natural Language Processing (NLP) is a subfield of artificial intelligence, linguistics and computer science that focuses on enablling computers to process natural language (Dialani 2020). One of the cases where NLP has been beneficial to people is where it has been used for machine translation, performing the task of translating from one language to another. In this case, NLP helps the computer or machine to attempt the conversion from one langauge to another. NLP can also assist in learning and prediction sentiment/opinion from sentences or text. This NLP capability is utilised by companies to understand how customers feel and their opinion about the company's products and services through the analysis of their social media posts and comments. Furthermore, the chatbots that are used in the customer services space are one of the examples of NLP application (Dialani 2020). Contextual chatbots and Virtual Text Assistant are now widely used but they mostly understand a limited number of languages, such as English. South African native languages do not have enough resources to be used to built such contextual Chatbots and Virtual Text Assistant. Therefore, the resources for native languages need to be created so that they can be used to build software agents that understand South African native languages (Duvenhage et al. 2017).

South Africa is a multilingual country with eleven langauges (two of which are European and nine are African languages); the African languages are Sepedi, Sesotho, Setswana, Siswati, Tshivenda, Xitsonga, isiZulu, isiNdebele and isiXhosa and on the other hand, European languages are English and Afrikaans. It is important to note that these languages are official in South Africa (Alexander 2021). In South Africa, we have a challenge with the nine African languages because they are resource-poor. There is a shortage of curated and annotated corpora to enable them to benefit from Natural Language Processing. Therefore, the purpose of this study is to focus specifically on the corpus creation and annotation for isiZulu and Siswati and perform

a topic classification tasks on the data.

2 Critical Natural Language Processing Components

Globalisation and the increase in digital communications have created the demand for NLP systems that enable fast communication between people speaking different languages. However, some languages are missing in these systems. For instance, there are roughly 7000 spoken languages on the planet and Most of them still are not included in the NLP systems, primarily because they do not have the labelled corpora to build those NLP systems (Baumann & Pierrehumbert 2014). These languages with scarce or no resources are low-resourced languages (Whyatt & Pavlović 2019). The language resources include (but are not limited to) the annotated corpora and core technologies. Examples of core technologies include lemmatisers, part of speech tagger and morphological decomposers (Eiselen & Puttkammer 2014). On the other hand, the languages with high resources are the ones that have most of the resources needed to build the NLP technology (Xu & Fung 2013).

The high-resourced languages include English, French, Finnish, Italian, German, Mandarin, Japanese, etc. (Bonab et al. 2019, Xu & Fung 2013) and low-resourced languages include languages such as isiZulu, isiXhosa, Siswati etc. (Bosch et al. 2008). A study, by Eiselen & Puttkammer (2014), focused on the low-resourced languages, namely, isiZulu and Siswati; stated that annotated corpora are one of the things that low-resourced languages lack. Thus, the isiZulu and Siswati datasets need to be annotated, as part of the process of making these languages accessible for NLP and by enriching these two languages. Hsueh et al. (2009) defines data annotation as the process of labelling the dataset(s), an important step when building machine learning models. Stenetorp et al. (2012) stated that manual data annotation is the most important, time-consuming, costly, and tedious task for NLP researchers. Therefore, automation tools are developed to perform these annotations.

The lack of curated and annotated data impede the process of fighting the shortage of resources for low-resourced languages in the NLP space (Niyongabo et al. 2020). Besides, established NLP methods often cannot be transferred on or to these languages without these corpora (Niyongabo et al. 2020). Niyongabo et al. (2020) collected the datasets of two closely related African languages - Kirundi and Kinyarwanda from two different sources. A total of 21268 and 4612 articles were annotated for Kinyarwanda and Kirundi respectively. The two datasets underwent a cleaning process that involved the removal of special language characters and stopwords. The sources were newspapers and websites. These datasets were annotated, based on the title and content of the contained articles, into the following categories: *Politics; Sport; Economy; Health, Entertainment; History; Technology; Tourism; Culture; Fashion; Religion, Environment; Education; and Relationship* (Rakholia & Saini 2016). Hence, a very similar task was performed in this work as part of language resources creation.

2.1 Data generation techniques for low-resourced languages

An existing approach utilised to mitigate the challenges of low-resourced language data, is the language translation approach. That is the low-resourced language gets translated into the resource-rich language (Tang et al. 2018). However, in most cases, this approach suffers from language biases and may be impractical to achieve in real life (Tang et al. 2018). Sometimes the direct translation may be impossible or inaccurate due to language differences. Hence, the translated data will require manual processing thereafter, which is tedious and time-consuming. Manually creating data for low-resourced languages is time-consuming but a good approach, moreover, it introduces minimal language biases and more accurate than translated datasets (Shamsfard n.d.).

Cross-lingual and transfer learning is one of the combinations of techniques frequently used or preferred in NLP due to its speed and efficiency

(Shamsfard n.d.). This further serves to highlight why all languages must have NLP resources such as annotated data to avoid data simulations that have unfavourable effects.

Data Augmentation is a method that generates a copy (or unique data) of the data by slightly altering the existing data (Duong & Nguyen-Thi 2021). It increases the size of small training data in ways that improve model performance (Abonizio & Junior 2020). Model performance is highly dependent on the quality and size of the training data. Data Augmentation addresses the issue of small training data that leads to the models losing their generalisability (Kobayashi 2018).

Work by Marivate et al. (2020) had a small data size of Sepedi and Setswana native languages, and incorporated word embeddings based-contextual augmentation to increase the dataset used to train classification models. Each training dataset was augmented 20 times while the test dataset remains unchanged. In their study, the new data created replaced the words (based on context) in the sentences. Hence a new sentence was formed as a result of applying Contextual Data Augmentation. Furthermore, Data Augmentation improved the performance of the classifiers (Marivate et al. 2020). In this current study, the same Data Augmentation (word embedding-based augmentation) was performed on the Siswati and isiZulu dataset to increase the data size.

2.2 Dealing with data imbalance

The Synthetic Minority Oversampling Technique (SMOTE) is another technique that can be adopted when the learning is done on an imbalanced dataset, since it solves the problem of class imbalance (Fernández et al. 2018). SMOTE works by generating synthetic examples through inserting different values(words) in minority class, the values are randomly picked from a defined neighbourhood within feature space . Minority class is selected, then obtain the k-nearest neighbours of the same minority class and therefore utilises the k- neighbours to create the new synthetic examples (Fernández et al. 2018).

2.3 Related work

Supervised learning models perform better on larger labelled datasets, which presents a challenge for low-resourced languages as they don't have enough data and annotating data can be expensive (Fang & Cohn 2017). Most prior studies focused on developing parallel corpora between low and resource-rich languages, but parallel corpora are often unavailable for some low-resourced languages (Fang & Cohn 2017). Work by Zoph et al. (2016) identified low-resourced languages and investigated the idea of distance learning on machine translation. Since English and French are resource-rich languages, the two languages trained a neural machine translation (NMT) (Zoph et al. 2016). An English-French neural machine translation (NMT) model was initially trained. Afterwards, the NMT model initialised another NMT model to be used on a low-resourced and high-resourced pair (e.g. Uzbek-English) (Zoph et al. 2016), as such utilising transfer learning. In this case, the low-resourced languages investigated for transfer were Uzbek, Hausa, Turkish and Urdu. The transfer learning was shown to improve the BLEU (bilingual evaluation understudy) for low-resourced Neural machine translation (Zoph et al. 2016).

Work by Nguyen & Chiang (2017) explored transfer learning between the two low-resourced languages Turkey and Uzbek by first pairing each language with English and then generating the parallel data. Then, split the words with Bytes Pair Encoding (BPE) to maximise the overlapping vocab (Nguyen & Chiang 2017). The model and word embedding are trained on the first language pair (Turkey-English) and then the same model parameters and word embeddings were transferred to the other model that trained the second language pair (Uzbek-English). This technique improved the BLEU by 4.3% (Nguyen & Chiang 2017).

The datasets of low-resourced South African languages, isiZulu collected from isolezwe and National Centre for Human Language Technology

(www.sadilar.org); and Sepedi collected from National Centre for Human Language Technology were used to evaluate the performance of open-vocabulary models on the small datasets, the evaluated models include n-grams, LSTM, RNN, FFNN, and transformers. The performance of the models was evaluated using the byte pair encoding (BPE). The RNN performed better than the rest of the models on both the isiZulu and Sepedi datasets (Mesham et al. 2021). Nyoni & Bassett (2021) explored the machine translation capability from the zero-short learning, transfer learning and multilingual learning on two South African languages, namely, isiZulu and isiXhosa; and one Zimbabwean language, that is Shona. The datasets were in language pair (parallel text), that is, English-to-Shona, English-to-Zulu, English-to-Xhosa and Zulu -to-Xhosa, with the pair English -to- Zulu being the target pair since it has the smallest datasets (sentence pair). The transfer learning and zero-short learning did not outperform the multilingual model which produced the Bleu score of 18.6 for the English-to-Zulu pair. Moreover, these results provide an avenue for the development and improvement of low resource translation techniques (Nyoni & Bassett 2021).

Work by Marivate et al. (2020) attempted to address the issue of lack of clear guidelines for low-resources languages in terms of collecting and curating the data for specific use in the Natural Language Processing domain. In their investigation, two datasets of news headlines written in Sepedi and Setswana were collected, curated, annotated, and fed into the machine learning classification models to perform text classification. The datasets were annotated by means of categorising the articles into the following categories based on context: *Legal; General News; Sports; Politics; Traffic News; Community Activities; Crime; Business; Foreign Affairs* (Marivate et al. 2020). The evaluation metric was the F1-score, which is a model performance measure. One of the models, Xgboost, performed well as compared to other models (Marivate et al. 2020).

3 Developing news classification models for isiZulu and Siswati languages

In this section we discuss data collection and cleaning processes together with the classification models building approach.

3.1 Data Collection, Cleaning and Annotation

We discuss the initial news data collection and annotation process. We further discuss the data collection process of the larger dataset that was used to build our word representations.

3.1.1 News data collection and annotation

The isiZulu news data was collected from Isolezwe, which is a Zulu-language local newspaper. The news articles published online on Isolezwe website (http://www.isolezwe.co.za) were scraped and stored in a csv file for further processing. The Siswati dataset (news headlines) was collected from the public broadcaster for South Africa, that is, SABC news LigwalagwalaFM Facebook page (https://www.facebook.com/ligwalagwalafm/). The Siswati data was also scraped and stored on a csv file. Lastly, to build word respresentations other isiZulu and Siswati datasets were collected from SADILAR (www.sadilar.org) and Leipzig Corpus(https://wortschatz.uni-leipzig.de) for the purpose of better generalising word representations. We collected 752 (full artilces and titles) in isiZulu and collected 80 Siswati news headlines.

Post data collection process, we worked to categorise the news items using the International Press Telecommunications Council (IPTC) News Categories (or codes)[1]. The categories used were: *1. disaster, accident and emergency incident, 2. economy, business and finance, 3. education, 4. environment, 5. health, 6. human interest, 7. labour, 8. lifestyle and leisure, 9. politics, 10. religion and belief, 11. science*

and technology, 12. society, 13. sport, 14. weather, 15. arts, culture, entertainment and media, 16. crime, law and justice, 17. conflict, war and peace. We make available the data, and annotations and data statement[2] [3]. An example of an annotated isiZulu article is shown below:

Politics
UMENGAMELI we-ANC uMnuz Cyril Ramaphosa ugqugquzele abantu basePort Shepstone nezindawo ezakhele leli dolobha ukuthi bagcwalise iMoses Mabhida Stadium lapho ezothula khona umyalezo wakhe weJanuary 8 ngoMgqibelo aphinde athule nomhlahlandela weqembu wokuheha abavoti njengoba kuyiwa okhethweni. URamaphosa ehambisana nabanye abaholi be-ANC esifundazweni uhambele kule ndawo izolo enxenxa abantu ukuthi batheleke ngobuningi kulo mgubho weqembu. Uphinde wathembisa ukuthi uzothula uhlelo lwakhe lokuthuthukisa izwe.

The isiZulu news (articles and titles) and Siswati news titles category distribution are shown below, it was observed that the datasets suffer from class imbalance, small data size and short text (only isiZulu and Siswati titles/headlines). Therefore, oversampling techniques, SMOTE and Data Augmentation were applied to mitigate class imbalance problem and also increase the data size.

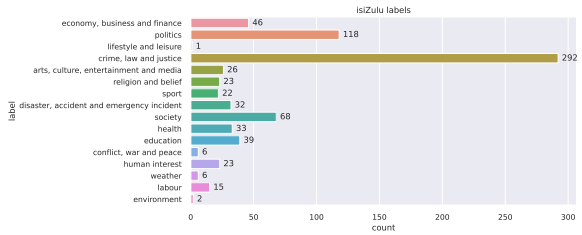

Figure 1: isiZulu initial Class Distribution

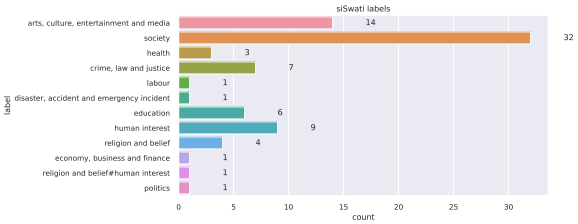

Figure 2: Siswati initial Class Distribution

For better modelling, class categories with few observations we revmoved, remaining with the below categories: *1. crime, law and justice, 2. economy, business and finance, 3. education, 4. politics, 5. society* for isiZulu and *1. crime, law and justice, 2. arts,culture,entertainment and media, 3. education, 4. human interest, 5. society* for Siswati. Since the number of class categories has dropped to 5 categories, the news dataset size also dropped to 563 (news articles and titles) for isiZulu and 68 (news titles) for Siswati The final datasets were cleaned and then used to build classification models, however, prior to model building, word representations were created using a larger datasets.

3.2 Data Preparation/Cleaning

All datasets collected in this work contained some noise such as single characters, white spaces, encoded characters, meaningless words, and special characters. The noise had to be removed before the datasets are fed into the models. All these noises on the datasets were removed. Below we explain each part of the followed cleaning step:

- The single characters carry less meaning, so they were removed from the datasets.
- There were instances where there are multiple spaces between two words, so those spaces were substituted with a single space.
- There were some characters/words that were not ASCII encoded then those characters were decoded back to ASCII.
- Special characters refer characters such as &%$ and they are not accepted by the models. Hence they were also removed.

- The data contained combination of letters that don't make any existing isiZulu/Siswati word. Words such as *'udkt','unksz','unkk'*

. Based on these criteria, they were also removed to streamline the corpus, and as a result, improve the analysis.

Since the datasets are noise-free, each letter in the datasets was set to lowercase, resulting in clean datasets to be used in machine learning models building.

3.2.1 Word Representations

It was stated above that the larger datasets collected from SADILAR and Leipzig Corpus for each language was used for word representations (vectorisers and embeddings) creation. The pre-trained vectorizers were created, enabling the opportunity to build classifiers with good generalisability in future. Therefore, from the collected corpora for each language, we created the following vectorizers: Bag Of Words, TFIDF and Word2vec (Mikolov et al. 2013).

Table 1: Vectorizer Corpora Sizes in number of tokens

	Tokens	
Source	**isiZulu**	**Siswati**
Sadilar	770845	399800
Leipzig	4296659	134827
Total	5067504	534627

3.3 News Classification Models

We arbitrarily selected a few classification algorithms to train models to perform news topic classification for isiZulu and Siswati datasets. The selected algorithm are Logistic Regression, XGBoost, Naive Bayes and LSTM.

We performed the classification on the original datasets, and then apply oversampling techniques, namely, Data Augmentation and SMOTE, to solve the class imbalance problem and increase the data size. The classification models were again executed on the Augmented and SMOTE datasets.

4 Experiments and Results

In this section we discuss the results obtained from the performed experiments, that is, the findings from the multiple combination of word representations and classification models on isiZulu and Siswati datasets. However, the findings presented here are basis, since this work only provide guidelines for resource creation of low-resource languages.

4.1 Experimental Setup

The maximum token size of 20 000 was used for both Bag Of Words and TFIDF vectorizers, whereas for Word2vec we used size 300. For each of the 4 classification models, 5-fold cross validation was applied during model training. As we are creating baseline models and working on small datasets (not enough to split into training, validation and test sets), then parameter optimisation was not performed in this work.

4.1.1 Baseline Experiments

In the baseline experiments, we train the classification models using 5-fold cross validation on isiZulu and Siswati original datasets and present the models performance for each dataset. The results show that Word2vec and LSTM model performed very well in all datasets as compared to other models. Below tables shows the classification model results obtained from original datasets.

Table 2: isiZulu Articles Original Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	21.73	21.12	16.34	52.4	(13.29,19.4)
Bag-Of-Words	Logistic Regression	41.23	34.97	36.06	54.53	(32.09,40.03)
Bag-Of-Words	XGBoost	49.14	31.33	32.51	54.89	(28.64,36.38)
TF-IDF	Naive Bayes	18.41	20.34	14.35	52.22	(11.45,17.24)
TF-IDF	Logistic Regression	32.09	26.13	24.19	54.71	(20.65,27.73)
TF-IDF	XGBoost	40.91	29.42	29.34	52.93	(25.58,33.1)
Word2vec	Naive Bayes	61.98	50.99	53.04	68.39	(48.91,57.16)
Word2vec	Logistic Regression	70.18	62.91	65.13	75.32	(61.19,69.07)
Word2vec	XGBoost	67.69	52.23	55.83	69.1	(51.73,59.93)
Word2vec	LSTM	**83.39**	**83.11**	**82.78**	**83.11**	(79.66,85.9)

Table 3: isiZulu Titles Original Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	17.6	20.62	15.33	51.69	(12.36,18.31)
Bag-Of-Words	Logistic Regression	18.36	21.83	17.38	52.76	(14.25,20.51)
Bag-Of-Words	XGBoost	20.91	21.23	17.03	51.51	(13.92,20.13)
TF-IDF	Naive Bayes	19.89	20.89	15.57	52.4	(12.57,18.56)
TF-IDF	Logistic Regression	20.47	21.9	17.58	52.93	(14.44,20.73)
TF-IDF	XGBoost	18.07	20.79	16.37	51.34	(13.31,19.43)
Word2vec	Naive Bayes	27.83	25.58	22.75	57.2	(19.29,26.22)
Word2vec	Logistic Regression	41.85	38.65	39.18	57.72	(35.14,43.21)
Word2vec	XGBoost	40.63	31.17	31.03	57.73	(27.21,34.85)
Word2vec	LSTM	**72.96**	**71.75**	**72.01**	**71.75**	(68.3,75.72)

Table 4: Siswati Titles Original Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	XGBoost	25.75	25.52	24.23	41.54	(14.05,34.42)
Bag-Of-Words	Naive Bayes	25.37	30	25.39	53.19	(15.04,35.73)
Bag-Of-Words	Logistic Regression	25.93	30.1	26.34	48.79	(15.87,36.81)
TF-IDF	Naive Bayes	13.61	22	15.61	48.68	(6.98,24.23)
TF-IDF	Logistic Regression	17.77	24	18.81	50.33	(9.52,28.1)
TF-IDF	XGBoost	25.16	29.33	25.5	47.58	(15.14,35.86)
Word2vec	Naive Bayes	31.77	34.76	31.57	59.01	(20.52,42.61)
Word2vec	Logistic Regression	29.59	32	28.09	57.58	(17.4,38.77)
Word2vec	XGBoost	28.77	31.43	27.96	54.84	(17.29,38.62)
Word2vec	LSTM	**87.53**	**80.88**	**81.06**	**80.88**	(71.75,90.37)

4.1.2 Augmentation

Data Augmentation is the technique that is used to increase the data size to improve the performance of the machine learning classifiers Oh et al. (2020). The most common way to augment the data is by means of replacing the words or phrases in a sentence by their synonyms where the synonym is derived by obtaining the semantically close words (Zhang et al. 2015).

The Siswati and isiZulu datasets were augmented using the same approach where the original words on the sentence are replaced based on their contextual meaning. The augmentation was done through referencing the words similarity from the Word2vec word embedding as per Marivate et al. (2020). Data Augmentation improved the performance of each model on all datasets as compared to original datasets, hence, it remains a task to investigate the effectiveness and robustness of this Data Augmentation algorithm, that can be achieved through comparing the algorithm results on resourced and low-resourced datasets.

The classification models trained on Word2vec outperformed all the classification models trained on TFIDF and Bag Of Words. For isiZulu articles, combination of Word2vec and XGBoost model outperformed all the models, scoring f1-score of 95.21%, on the other hand, Word2vec and Logistic Regression model combination performed well on isiZulu titles dataset scoring f1-score of 86.42%. Lastly, Word2vec and LSTM model combination performed well on Siswati titles dataset scoring f1-score of 93.15%. It was observed that iziZulu articles dataset scored high f1-score as compared to isiZulu titles, which explains that long texts improves the classification accuracy, and also highlights that Logistic Regression outperforms XGBoost on short text dataset. It remains a task to run the same comparison on Siswati dataset, as it was not covered in this work due to lack of Siswati full news articles dataset.

Table 5: isiZulu Articles Augmented Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	71.65	68.55	68.42	68.89	(65.87,70.97)
Bag-Of-Words	Logistic Regression	83.35	83.92	83.09	83.23	(81.04,85.15)
Bag-Of-Words	XGBoost	74.28	73.85	73.68	73.51	(71.26,76.09)
TF-IDF	Naive Bayes	75.71	73.77	73.6	73.98	(71.18,76.02)
TF-IDF	Logistic Regression	79.65	79.91	79.2	79.39	(76.97,81.42)
TF-IDF	XGBoost	80.44	80.44	79.92	80.02	(77.72,82.11)
Word2vec	Naive Bayes	72.37	71.79	71.79	71.31	(69.32,74.26)
Word2vec	Logistic Regression	91.6	91.9	91.3	91.3	(89.75,92.84)
Word2vec	XGBoost	95.54	**95.73**	**95.21**	**95.14**	(94.04,96.39)
Word2vec	LSTM	**96.08**	94.45	94.45	94.45	(93.2,95.71)

Table 6: isiZulu Titles Augmented Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	58.93	32.91	31.62	37.83	(28.86,34.37)
Bag-Of-Words	Logistic Regression	60.79	34.54	34.05	39.2	(31.24,36.85)
Bag-Of-Words	XGBoost	51.12	28.22	24.47	33.27	(21.92,27.01)
TF-IDF	Naive Bayes	59.45	33.25	32.3	38.1	(29.54,35.07)
TF-IDF	Logistic Regression	59.41	34.87	34.42	39.47	(31.6,37.23)
TF-IDF	XGBoost	53.33	28.85	25.41	33.82	(22.83,27.98)
Word2vec	Naive Bayes	67.92	57.97	59.3	60.89	(56.39,62.21)
Word2vec	Logistic Regression	**86.35**	**87.65**	**86.42**	**85.69**	(84.39,88.45)
Word2vec	XGBoost	86.2	85.99	85.83	84.96	(83.77,87.89)
Word2vec	LSTM	85.32	85.16	84.37	85.16	(82.22,86.52)

Table 7: Siswati Titles Augmented Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	71.98	69.52	69.35	68.79	(61.82,76.88)
Bag-Of-Words	Logistic Regression	78.78	74.8	74.74	74.31	(67.65,81.84)
Bag-Of-Words	XGBoost	81.99	74.7	74.47	74.33	(67.35,81.59)
TF-IDF	Naive Bayes	75.67	73.03	72.85	72.24	(65.58,80.11)
TF-IDF	Logistic Regression	78.93	75.5	75.57	75	(68.55,82.59)
TF-IDF	XGBoost	81.1	74.13	73.09	73.62	(65.84,80.33)
Word2vec	Naive Bayes	84.26	83.41	82.52	82.66	(76.32,88.73)
Word2vec	Logistic Regression	91.17	89.9	87.83	88.89	(82.49,93.17)
Word2vec	XGBoost	91.57	91.33	89.8	90.22	(84.86,94.74)
Word2vec	LSTM	**94.88**	**92.41**	**93.15**	**92.41**	(89.02,97.27)

4.1.3 SMOTE

SMOTE is an oversampling technique used to rebalance the original training set through the creation of synthetic samples of the minority class Fernández et al. (2018). This technique works by selecting the minority class and the total amount of oversampling to balance the classes, then the k-nearest neighbours for that particular class are obtained , therefore, iteratively the k nearest neighbours are randomly chosen to create new instances Fernández et al. (2018). This oversampling technique was used to balance the classes and increase the dataset. Note that SMOTE uses a different approach from the Data Augmentation approach presented earlier.

We applied SMOTE on our three datasets and run the classification model using 5-fold cross validation, the results from each dataset are presented below. From the below tables, it was observed that Word2vec produced the best classification models from all the three datasets. XGBoost performed well in all instances scoring f1-score of 93.35%, 91.26%, 87.46% for isiZulu articles, isiZulu titles and Siswati titles datasets respectively. We observed the XGBoost model on isiZulu articles struggled to separate *society* and *politics* from *crime,law and justice* since most of the incorrect classification happened in the instance where *society* and *politics* were classified as *crime,law and justice*.

Table 8: isiZulu Articles SMOTE Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	36.37	39.06	36.63	39.04	(34.16,39.11)
Bag-Of-Words	Logistic Regression	55.67	51.19	50.08	51.16	(47.52,52.65)
Bag-Of-Words	XGBoost	82.31	76.34	75.99	76.37	(73.8,78.18)
TF-IDF	Naive Bayes	78.93	77.81	76.83	77.81	(74.67,79.0)
TF-IDF	Logistic Regression	82.2	82.38	81.68	82.4	(79.7,83.66)
TF-IDF	XGBoost	81.7	79.17	79.51	79.18	(77.44,81.58)
Word2vec	Naive Bayes	74.44	74.25	74.12	74.25	(71.87,76.37)
Word2vec	Logistic Regression	92.43	92.11	91.88	92.12	(90.48,93.28)
Word2vec	XGBoost	**93.75**	**93.55**	**93.35**	**93.56**	(92.08,94.63)

Table 9: isiZulu Titles SMOTE Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	36.92	23.33	15.91	23.22	(14.03,17.78)
Bag-Of-Words	Logistic Regression	46.08	25.9	18.23	25.89	(16.25,20.21)
Bag-Of-Words	XGBoost	65.14	38.34	37.52	38.36	(35.03,40.0)
TF-IDF	Naive Bayes	64.37	37.69	37.38	37.6	(34.9,39.86)
TF-IDF	Logistic Regression	65.23	39.72	39.71	39.73	(37.2,42.22)
TF-IDF	XGBoost	65.6	38.2	37.56	38.22	(35.08,40.05)
Word2vec	Naive Bayes	74.49	74.02	73.85	74.04	(71.6,76.11)
Word2vec	Logistic Regression	91.56	91.08	90.63	91.1	(89.13,92.12)
Word2vec	XGBoost	91.96	91.56	91.26	91.58	(89.81,92.71)
Word2vec	LSTM	73.53	72.82	72.75	72.82	(69.08,76.43)

Table 10: Siswati Titles SMOTE Dataset Model Performance

Preprocessing	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
Bag-Of-Words	Naive Bayes	60.63	40.67	37.91	40	(30.4,45.43)
Bag-Of-Words	Logistic Regression	65.03	44.19	42.91	44.38	(35.24,50.58)
Bag-Of-Words	XGBoost	81.3	74.38	73.65	74.38	(66.83,80.48)
TF-IDF	Naive Bayes	80.71	79.14	74.32	78.75	(67.55,81.09)
TF-IDF	Logistic Regression	82.25	82.95	80.42	83.12	(74.27,86.57)
TF-IDF	XGBoost	85.47	77.05	76.6	76.88	(70.04,83.16)
Word2vec	Naive Bayes	85.86	83.71	82.5	83.75	(76.62,88.39)
Word2vec	Logistic Regression	**90.35**	88.1	86.2	88.12	(80.86,91.55)
Word2vec	XGBoost	89.88	**88.76**	**87.46**	**88.75**	(82.33,92.59)

5 Summary

We observed that Data Augmentation outperformed SMOTE in two instances, that is, isiZulu articles and Siswati titles datasets, whereas SMOTE outperformed Data augmentation only in case of isiZulu titles dataset, however, we hope to look into the difference performance from these re-sampling techniques and have a confirmatory pipeline to provide guidance on what approach to take under what circumstance. However, we present the generalised pipeline obtained from this work as a baseline.

The Pipeline obtained from this work was summarised and presented in figure 3 below together with the corresponding top performing classification models presented in table 3, the figure 3 shows the choices that produced the best results under different circumstances for three different datasets. It was observed that the datasets used resembled three different qualities, that is, large size and long-text (isiZulu Articles), large size and short text(isiZulu Titles), and small size and short text(Siswati), these varieties produced different outcomes from the models under the same circumstance and can be generalised as follows:

- If the data size is large and contains long-text then Contextual Data Augmentation is recommended over SMOTE, and LSTM is likely to perform better.

- If the data size is large and contains short-text then SMOTE is recommended over Contextual Data Augmentation, and XGBoost is likely to perform better.
- If the data size is small and contains short-text then Contextual Data Augmentation is recommended over SMOTE, and XGBoost is likely to perform better

The Above generalisation is limited to Word2vec word embedding since it is the one that produced outstanding results from all the datasets as compared to TFIDF and Bag-Of-Words. It remains a task to further investigate the poor performance from TFIDF and Bag-Of-Words, possibly the parameter change in classification could lead to good results.

Table 11: Top Performing Classification Models

Best Model based on Sampling technique								
Dataset	Sampling	Word embbeding	Model	Precision(%)	Recall(%)	F1-score(%)	Accuracy(%)	Confidence Interval(f1 score)
isiZulu Articles	Augmented	Word2vec	XGBoost	95.54	95.73	95.21	95.14	(94.04,96.39)
isiZulu Titles	Augmented	Word2vec	Logistic Regression	86.35	87.65	86.42	85.69	(84.39,88.45)
Siswati Titles	Augmented	Word2vec	LSTM	94.88	92.41	93.15	92.41	(89.02,97.27)
isiZulu Articles	SMOTE	Word2vec	XGBoost	93.73	93.55	93.35	93.56	(92.08,94.63)
isiZulu Titles	SMOTE	Word2vec	XGBoost	91.96	91.56	91.26	91.58	(89.81,92.71)
Siswati Titles	SMOTE	Word2vec	XGBoost	89.88	88.76	87.46	88.75	(82.33,92.59)

6 Conclusion and Future Work

This work introduced the collection and annotation of isiZulu and Siswati news datasets. There is still a data shortage (more especially annotated data) of these two native languages, especially Siswati. However, this work paved a way for the other researchers who would want to use annotated data for isiZulu and/or Siswati in downstream NLP tasks.

The experimental findings from the classification models and different combinations of word embeddings with model baselines were presented. Though we were limited by the data availability, however, this provides an overview of what could be achieved with minimal datasets. The isiZulu and Siswati annotated datasets will be made available for other researchers, the pre-trained vectorizers will be open-sourced to other researchers and the classification

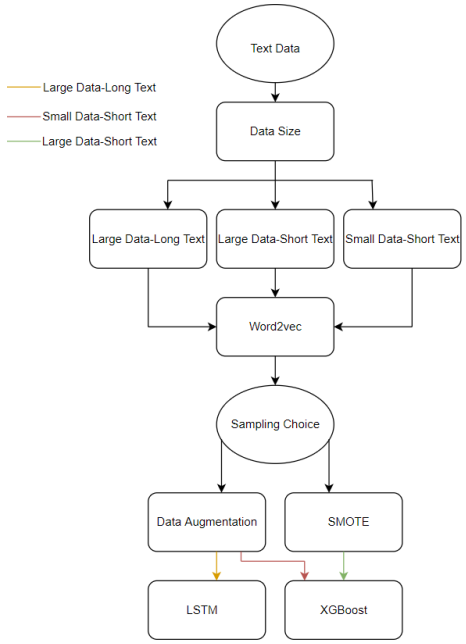

Figure 3: Recommended Pipeline

results that maybe be used as benchmarks.

The collection and annotation of native language datasets remain a task for the future. For this to be successful, there needs to be an identification of other language sources where the dataset can be extracted for more models to be trained. Furthermore, NLP researchers need to focus more on effective ways to augment the datasets. They should be compared with SMOTE sampling, because of the imbalance in the dataset. It is beneficial to have effective ways to augment native language datasets.

In addition, it is also worth investigating the poor performance of TFIDF and Bag-Of-Words compared to Word2vec, possible investigation areas could be the word embedding nature and the classification models hyperparameters optimisation that could improve classification performance. Another extension of this work is transfer learning from isiZulu to Siswati. The isiZulu dataset is large compared to the Siswati dataset making it a viable avenue of research to investigate if transfer learning improves the classification performance for Siswati in this context.

Notes

[1] https://iptc.org/standards/newscodes/

[2] https://github.com/dsfsi/za-isizulu-siswati-news-2022

[3] https://doi.org/10.5281/zenodo.7193346

References

Abonizio, H. Q. & Junior, S. B. (2020), Pre-trained data augmentation for text classification, *in* 'Brazilian Conference on Intelligent Systems', Springer, pp. 551–565.

Alexander, M. (2021), 'The 11 languages of south africa'.
URL: *https://southafrica-info.com/arts-culture/11-languages-south-africa/*

Baumann, P. & Pierrehumbert, J. (2014), Using resource-rich languages to improve morphological analysis of under-resourced languages, *in* 'Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)', pp. 3355–3359.

Bonab, H., Allan, J. & Sitaraman, R. (2019), Simulating clir translation resource scarcity using high-resource languages, *in* 'Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval', pp. 129–136.

Bosch, S., Pretorius, L. & Fleisch, A. (2008), 'Experimental bootstrapping of morphological analysers for nguni languages', *Nordic Journal of African Studies* **17**(2), 23–23.

Dialani, P. (2020), 'What is nlp and why is it important?'.
URL: *https://www.analyticsinsight.net/what-is-nlp-and-why-is-it-important/#:~:text=Natural language%2*

Duong, H.-T. & Nguyen-Thi, T.-A. (2021), 'A review: preprocessing techniques and data augmentation for sentiment analysis', *Computational Social Networks* **8**(1), 1–16.

Duvenhage, B., Ntini, M. & Ramonyai, P. (2017), Improved text language identification for the south african languages, *in* '2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)', IEEE, pp. 214–218.

Eiselen, R. & Puttkammer, M. J. (2014), Developing text resources for ten south african languages., *in* 'LREC', pp. 3698–3703.

Fang, M. & Cohn, T. (2017), 'Model transfer for tagging low-resource languages using a bilingual dictionary', *arXiv preprint arXiv:1705.00424* .

Fernández, A., Garcia, S., Herrera, F. & Chawla, N. V. (2018), 'Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary', *Journal of artificial intelligence research* **61**, 863–905.

Hsueh, P.-Y., Melville, P. & Sindhwani, V. (2009), Data quality from crowdsourcing: a study of annotation selection criteria, *in* 'Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing', pp. 27–35.

Kobayashi, S. (2018), 'Contextual augmentation: Data augmentation by words with paradigmatic relations', *arXiv preprint arXiv:1805.06201* .

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R. & Modupe, A. (2020), 'Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi', *arXiv preprint arXiv:2003.04986* .

Mesham, S., Hayward, L., Shapiro, J. & Buys, J. (2021), 'Low-resource language modelling of south african languages', *arXiv preprint arXiv:2104.00772* .

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Weinberger, eds, 'Advances in Neural Information Processing Systems',

Vol. 26, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf*

Nguyen, T. Q. & Chiang, D. (2017), 'Transfer learning across low-resource, related languages for neural machine translation', *arXiv preprint arXiv:1708.09803* .

Niyongabo, R. A., Qu, H., Kreutzer, J. & Huang, L. (2020), 'Kinnews and kirnews: Benchmarking cross-lingual text classification for kinyarwanda and kirundi', *arXiv preprint arXiv:2010.12174* .

Nyoni, E. & Bassett, B. A. (2021), 'Low-resource neural machine translation for southern african languages', *arXiv preprint arXiv:2104.00366* .

Oh, C., Han, S. & Jeong, J. (2020), 'Time-series data augmentation based on interpolation', *Procedia Computer Science* **175**, 64–71.

Rakholia, R. M. & Saini, J. R. (2016), Lexical classes based stop words categorization for gujarati language, *in* '2016 2nd international conference on advances in computing, communication, & automation (ICACCA)(Fall)', IEEE, pp. 1–5.

Shamsfard, M. (n.d.), 'Challenges and opportunities in processing low resource languages: A study on persian'.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S. & Tsujii, J. (2012), Brat: a web-based tool for nlp-assisted text annotation, *in* 'Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics', pp. 102–107.

Tang, X., Cheng, S., Do, L., Min, Z., Ji, F., Yu, H., Zhang, J. & Chen, H. (2018), 'Improving multilingual semantic textual similarity with shared sentence encoder for low-resource languages', *arXiv preprint arXiv:1810.08740* .

Whyatt, B. & Pavlović, N. (2019), 'Languages of low diffusion and low resources: translation research and training challenges: Special issue proposal itt–15 (1), march 2021'.

Xu, P. & Fung, P. (2013), 'Cross-lingual language modeling for low-resource speech recognition', *IEEE transactions on audio, speech, and language processing* **21**(6), 1134–1144.

Zhang, X., Zhao, J. & LeCun, Y. (2015), 'Character-level convolutional networks for text classification', *Advances in neural information processing systems* **28**, 649–657.

Zoph, B., Yuret, D., May, J. & Knight, K. (2016), 'Transfer learning for low-resource neural machine translation', *arXiv preprint arXiv:1604.02201* .

How to deal with so-called "*vowel verbs with variant non-vowel forms*" in a bilingual Zulu lexicon ? The case for a pragmatic approach.

Michel Lafon[1]
research fellow (retired),
CNRS Llacan Paris, IFAS - Johannesburg
& CenterPoL, University of Pretoria,
maikoro12@gmail.com
& Bolofo Mongezi
PhD Candidate, Wits School of Education

Abstract

Whilst lemmatisation in Zulu (and cognate languages) remains to this day a partially unsettled issue, with dictionaries, both bilingual and monolingual, adopting differing strategies in respect of nouns, we wish to focus here on a somewhat minute aspect of the matter: how to lemmatize verbs with optional initial vowels. As the variation occurs at the initial, it is crucial that it receives proper dictionary treatment, lest the user be misdirected or misinformed.

Keywords: isiZulu - lexicography - vowel verbs

We shall start by presenting verbs with optional initial vowels in Zulu, contrasting the lemmatization strategies offered by main available dictionnaries, some of which in our view lack consistency. We shall then introduce to the pragmatic solution we adopted in our bi-directional French to Zulu and Zulu to French lexicon, which we claim gives better justice to the language.

1 The corpus

Zulu (or isiZulu)[2] is in terms of demography the main African (Bantu) language spoken in South-Africa,[3] being in its different sociolects the first or home language of close to 13 millions people, viz., 22% of a population of at least 59 millions.[4] Although there exists a number of regional or sub-regional dialects, its standard form, taught in schools and propagated through formal situations, including literature, the written and spoken press, as well as official ceremonies, is widely accepted. It also boasts a well-accepted orthography. This paper is based on the standard language and our data is taken mainly from secondary sources, *id est*, dictionaries as quoted in the paper, complemented by random observations taken from formal and informal written documents as well as informal speech, and the native-speaker fluency of one co-author.

Zulu has a rather large inventory of what (Clement Martyn Doke 1992, 131) refers to as "*vowel verbs with variant non-vowel forms*" , such as

> **-edlula** or **-dlula,**[5] *pass*[6]
>
> **-esutha** or **-sutha,** *be satiated with food*

The initial vowel is mostly /e/, but /a/ and /o/ also occur, more often as further variants (examples from dictionaries) :

> **-ehlula, -ahlula** or **-hlula,** *conquer, defeat*
>
> **-ejwayela, -ojwayela** or **—jwayela,** *be used to*

-ephula, -aphula, -ophula, or **—phula,** *break*

Whereas the presence in the standard of variant forms may proceed from the incorporation of regional features as, although based mainly on « *the Central Zululand Dialect* » (Kubeka 1979, 83), it includes arguably features of other varieties, it seems that the variation has now become evolutionary, as some of Doke's vowel-initial verbs do not seem valid today (but see further down):

?-efunda for **-funda** *study*

?-egcwala for **-gcwala,** *be full*

?-emithi for **-mithi** *be with small (animals); become pregnant*

?-ojwayela for **-jwayela,** *be used to*

This would point to a phenomenon of vowel erosion. However the fact that some of those verbs, like **-gcwala**, seem to have their source in ideophones (**gcwa**, *of being full to the brim)* which are consonant-commencing suggests this explanation is not sufficient. The vowel would then have been added (or the ideophone would have been obtained by deletion of the vowel). We shall not investigate this issue further however, as it is not relevant to the problem at hand.

Nouns derived from the same root show the same alternative :

abalusi or **abelusi,** *herders* (cf. **—elusa** or **-lusa,** *herd)*

but all possibilities are not always admitted:

umelaphi[7] *healer* **: ? umlaphi**,

The same variation may apply to extended verbs and nouns derived from them:

-ehluleka or **-ahluleka,** *fail* and **isehluleki** or **isahluleki** *failure*

-efundisa or **-fundisa** *teach,* and a**bafundisi** or **abefundisi,** *priests*

-efundisa seems clearly obsolete but the noun **abefundisi** is attested

It also happens that a variant ∂ does not cover all meanings of variant ß, suggesting that the verbs were originally different but became conflated due to the phonetic loss of the initial vowel:

-phuza has now two meanings: **—phuza,** *drink* and **-phuza** (variant of **-ephuza)** *be late*

To this category we add the three (or four) verbs with /i/ as the initial vowel, that Doke in the passage quoted above refers to as "*latent-vowel verbs*" due to the fact that /i/ may only appear in conjugated forms when preceded by a vowel /a/ with which it then coalesces resulting in /e/:[8]

-(i)ma: *stand:* **ngiyema** or **ngiyama,** *I stand*

-(i)-za, *come:* **bayeza** or **bayaza,** *they are coming*

-(i)zwa, *understand:* **ubezwa** or **ubazwa,** *listen to them*

Since in Zulu /i/ tends to disappear when following any vowel other than /a/ (or merge if the preceding vowel is /i/), no conclusion can be drawn from instances where initial /i/ does not surface, as to what is the variant of the verb stem in any given instance:

ngimile, *I stood*: **lngil-liml-lilel** or **lngil-lml-lilel**

uzwile, *you heard*: **lul-lizwl-lilel** or **lul-lzwl-lilel**

We therefore posit for those verbs a variant form of the stem with an initial vowel, akin to the situation above :

-ima or **-ma,** *stand*

-iza or **-za,** *come*

-izwa or **-zwa,** *hear*, *understand*

The choice of variant appears related to the tense form, the presence of an extension after the root which seems to result in a preference for the shorter, consonant-commencing form, as well as to the stem itself. The following exemples are drawn from our own observations of unsollicited speech:

ngiyezwa rather than **ngiyazwa,** *I hear*

ngiyazwisisa rather than **ngiyezwisisa,** *I understand very clearly*

kuyezwakala or **kuyazwakala,** *it is understandable*

ngiyeza rather than **ngiyaza,** *I am coming*

ngiyabazisa endlini rather than **ngiyabezisa endlini,** *I make them come home*

ngiyema or **ngiyama,** *I stand*

ngiyamela abafazi rather than **ngiyemela abafazi,** *I am waiting for the women*

2 Treatment in Dictionaries

How are these facts dealt with in major Zulu dictionaries?

Keeping to items already mentioned, we compare the main dictionaries presently available:[9] bilingual - Doke & al. Zulu-English and English-Zulu (Clement M. Doke et al. 1999), Dent and Nyembezi's Scholar's Zulu (Dent and Nyembezi 1995), de Schryver's (de Schryver 2015)- and monolingual - Nyembezi's (Nyembezi 1992) and Mbatha's (Mbatha 2006). We are aware that these dictionaries were not compiled within the same framework. Whereas the four older follow with small deviations between them the strategy set out by Doke for Zulu lexicography, which lemmatizes stems rather than words, de Schryver's, besides being corpus-based, introduces an approach which seeks to lemmatize words rather than stems (de Schryver and Wilkes 2008). However, this revolution in Zulu lemmatization does not really affect verbs. Even though Doke and Nyembezi quote verbs under the imperative while other dictionaries refer to the verb stem, appropriately defined in (Marlo 2013) as "*an obligatory root, one or more possibly occurring derivational suffixes* [also known as extensions]*[10] and an inflectional final suffix commonly called the "Final Vowel"*, the form is segmentally similar, the singular imperative being none other in Zulu than the said stem with a specific tone structure.

We use the following symbols:

A : main entry (as revealed by length of description);

A': main entry although shorter than A [that implies two main entries in same dictionary]

A + a : main entry + cross-reference to at least one variant;

B : shortened entry, with cross-reference to main;

B' : cross-reference to main ;

0 : no mention of the item

	-phula	**-ephula**	**-aphula**	**-esutha**	**-sutha**	**-egcwla**	**-gcwala**
	break			*be satiated*		*become full*	
Doke	B	B	A + a	A′	A	B	A + a
Dent	A	A	A	A′	A	0	A
de Schryver	A	0	0	0	0[11]	0	A
Nyembezi	A′	A	A′	A′	A	0	A
Mbatha	A	A′	A′	A	A	0	A

	-dlula	**-edlula**	**-hlula**	**-ehlula**	**-ahlula**	**-emitha**[12]	**-mitha**
	pass		*conquer*			*become pregnant*	
Doke	A + a	B	B	B	A +a	a	A + a
Dent	A	B′	A	A	A	A	0
de Schryver	B′	A	B′	A	B′[13]	0	A
Nyembezi	A	A′	A	A′	A′	A′	A
Mbatha	A	A′	A	A′	A′	0	A

	-ima	**-ma**	**-iza**	**-za**	**-imba**	**-mba**	**-izwa**	**-zwa**
	stand		*come*		*dig*		*hear*	
Doke	0	A + a	0	A + a	0	A + a	0	A + a
Dent	0	A	0	A	0	0	0	A
de Schryver	0	A		A	0	0[12]	A	0
Nyembezi	0	A	0	A	0	A	0	A
Mbatha	0	A	0	A	0	A	0	A

3 Observations

It seems, even judging from such a limited and haphazard sample, that inconsistency prevails within dictionaries as well as between them, as a cursory look at the table suggests.

- Zulu to English.

Since Doke refers specifically to "*vowel verbs with variant non-vowel forms*" one would expect these verbs to be entered systematically under the vowel with an indication that a consonant-commencing variant exists : not so but most instances include a cross-reference, which mitigates the issue. In Dent on the other hand repetition is frequent: **-ephula**, **-aphula** and **-phula** among other such examples constitute three different entries with almost the same definitions repeated without any cross-referencing. De Schryver is more consistent in cross-referencing the (supposedly) less frequent form to the main one.

Regarding the so-called latent vowel or i-commencing verbs, Doke is the only one to systematically make mention of "*latent i*" in the description of all corresponding entries. Dent and de Schryver make no explicit allusion to it while implying it nevertheless through examples: under **-zwa**, Dent gives **ukungezwa nakutshelwa**, *to want to see by oneself* ("*not wanting to be told*"), where the negative marker **-nga-** alters to **nge** as its /a/ vowel coalesces with the initial /i/;[14][15]. In a similar fashion, under **-za** de Schryver gives **abantu beza ngobuningi**, *people came in their numbers*, whereas **beza** cannot be obtained from the subject prefix **ba-** and the verb given as **-za**. Same situation under **zwakala** (separate entry from **-zwa**) with **kuyezwakala** *it is understandable*.[16]

- English to Zulu

In all three bilingual dictionaries, only the form given as main entry in the Zulu to English is indicated, with no mention whatsoever of variants. That applies *inter alia* to Dent where, under *break*, only **-aphula** is listed.

- Zulu monolingual

Nyembezi and Mbatha provide almost similar descriptions for each variant, each treated as a main entry (see **-dlula**, **-edlula** *inter alia*), occasionally indicating the existence of a variant assimilated to a synonym like in the **-hlula** and **-phula** series. Even if variants are somehow synonyms, it would make sense in a linguistic work to discriminate. And that does not preclude inconsistencies: one of Mbatha's examples under **dlula** is **ukwedlula ngendlu yakhiwa** *not to offer assistance to people working*, which should appear under **-edlula**. Same for **uwephulile umoya wami lo mfana**, *this boy broke my heart,* which should illustrate **ephula** rather than **-phula**. As for the "latent vowel", no explicit mention whatsoever. However Mbatha includes in his examples conjugated forms where the "latent i" does appear: under **-za**: (…) **uyeza**, (…) *you are coming*; under **-zwa**: (…) **sengathi akezwa** (…) *as if he does not understand*; **ukungezwa ngokutshelwa** (see above and note 14).

4 A pragmatic approach

Our bidirectional French / Zulu lexicon (Lafon and Mongezi 2022) has no pretention to be extensive but rather a handy support whilst one is engaged in conversation. Space and cost were huge concerns. Still we aimed at covering as much vocabulary as possible whilst attempting to remain consistent throughout.

In the French to Zulu part, we opted to place the possible vowel within brackets, so as to offer a maximum of information in a minimum of space, treating i-commencing verbs in a similar fashion:

passer (to pass): **-(e)dlula ;**
aller (to come): **-(i)za**

Thus all variants of conjugated forms can be deducted from the lemmas as they stand.

Obviously the same strategy could not be followed in the Zulu to French part. In order to avoid repetition as well as omission, we opted to lemmatize the form which appeared to us the more common (based on native speaker's intuition), the other or possibly other form(s) being however systematically listed and cross-referred to the one chosen as lemma :

-edlula : voir [see] **-dlula**
-iza : voir [see] **-za**

This strategy counterbalanced at least partially any wrong estimate on our side of the relative frequency of variants. The strategy obviously is not ignored by all of the dictionaries reviewed;[17] rather it is not applied systematically.

5 Conclusion

It would seem, from the review of lexicographical treatment of a few items in old as well as recent Zulu dictionaries, both bilingual and monolingual, traditional as well as recent and corpus- and frequency-based (de Schryver's), that the twin issues of "*vowel verbs with variant non-vowel forms*" and "*latent-vowel verbs*" have been largely overlooked. Minor as this may seem, it remains a blind spot. Our *ad hoc* solution, to consistently include all variants in the simplest and less cumbersome manner, might then go some way towards achieving better coverage in that particular instance. This reminds us that there is no Holy Grail to achieving consistency and user-friendliness in the lexicography of Zulu and cognate languages. Lexicographical issues have to be considered taking into account their specifics, this being the only manner to provide adequate, complete and easily accessible information.

6 References

Bryant, Alfred T. 1905. A Zulu-English Dictionary: With Notes on Pronunciation, a Revised Orthography and Derivations and Cognate Words from Many Languages; Including Also a Vocabulary of Hlonipa Words, Tribal-Names. Etc., a Synopsis of Zulu Grammar and a Concise History of the Zulu People from the Most Ancient Times. Maritzburg: Davis & Sons.

Colenso, John William. 1861. Zulu English Dictionary. Vol. 1. Pietermaritzburg.

Dent, G. R., and C. L. Sibusiso Nyembezi. 1995. Scholar's Zulu Dictionary; English-Zulu, Zulu-English. [1969]. Pietermaritzburg: Shuter and Shooter.

Doke, Clement M., D. Mc K. Malcolm, J. M. A. Sikakana, and B. Wallet Vilakazi. 1999.

English-Zulu, Zulu-English Dictionary. [1958]. Wits University Press.
Doke, Clement Martyn. 1992. Textbook of Zulu Grammar. [1927]. 1 vols. Cape Town: Longmans Southern Africa.
Hadebe, Samukele, ed. 2001. Isichazamazwi SesiNdebele. Vol. 1. Harare: College Press Publishers and The African Language Research Institute (ALLEX).
Kubeka, Isaac Sibusiso. 1979. 'A Preliminary Survey of Zulu Dialects in Natal and Zululand'. Univ of Natal, dep of Zulu language & literature.
Lafon, Michel, and Bolofo Mongezi. 2022. Lexique Français-Zoulou; Isichayamagama SesiZulu Kuya KwisiFulentshi. Moroni: KomEdit.
Marlo, Michael R. 2013. 'Verb Tone in Bantu Languages: Micro-Typological Patterns and Research Models'. Africana Linguistica XIX: 137–234.
Mbatha, M. O. 2006. Isichamazwi SesiZulu. Pietermaritzburg, South Africa.
Nyembezi, C. L. Sibusiso. 1992. AZ : Isichazimazwi Sanamuhla Nangomuso. Pietermaritzburg: Reach Out Publishers.
Samuelson, R.C.A. 1923. The King Cetywayo Zulu Dictionary. Durban.
Schryver, Gilles-Maurice de, ed. 2015. Funda IsiNgisi. Learn IsiZulu. IsiZulu-IsiNgisi IsiNgisi-IsiZulu. Isichazamazwi Sesikole. School Dictionary. 2nde ed. Cape Town: Oxford University Press.
Schryver, Gilles-Maurice de, and Arnett Wilkes. 2008. 'User-Friendly Dictionaries for Zulu: An Exercise in Complexicography'. In Proceedings, edited by Elisenda Bernal and Janet DeCesaris, 827–36. Barcelone, Institut Universitari de Linguistica Applicada.

[1] Corresponding author

[2] It is considered appropriate in South Africa to refer to local African (Bantu) languages with their prefixes in the name of respecting self-identification. We have however opted to use consistently English language-naming conventions. Thus we maintain French rather than « français » in the text.

[3] Bantu is a convenient classificatory term for a family of languages indigenous to the African continent. The term has no implication outside the linguistic field. Languages of this family share a number of features, including a noun class system characterized by prefixes and agreement schemes.

[4] Extrapolated from the 2011 Census which included a question on ethnicity, and the population figures for 2020 (see https://www.statssa.gov.za). This figure probably does not give justice to recent immigration. The results of the 2022 census are not published to date.

[5] Verbs are quoted under their stem, preceded by a hyphen to indicate that any conjugated form - apart from the imperative- requires prepositioned elements. See also below.

[6] Translations are given so as to facilitate identification of each item but by no means claim to cover the whole shade of meaning of any given word or stem.

[7] Heard in one episode of the Zulu-spoken soapie **Uzalo** 207, November 2018, SABC 1

[8] Doke includes **-iva**, *increase*, but we keep it out due to its obsolescence. **-ipha** and **-ikha** are sometimes given as variants for **-pha** *give* and **-kha** *draw water* respectively, but these not being commonly accepted, we refrain from considering them either.

[9] We did not have access to previous substantial works, such as (Colenso 1861), (Samuelson 1923) or (Bryant 1905), by his own admission main sources of Doke.

[10] Our addition.

[11] Item apparently missed or omitted by de Schryver, which is puzzling as it does seem a frequent enough term with forms such as **ngisuthi,** *I am satiated*, etc

[12] **-emithi** and **-mithi** in Doke.

[13] No specific entry *-ahlula but passive **-ahlulwa** refered to **-ehlulwa**

[14] Cf. for instance **ukungavumi ukutshelwa**, *not accepting to be told* where **-nga-** remains unaffected as it precedes a consonant.

[15] As Dent's verb examples keep to the verbal stem, one is hard pressed to see if instances of "latent i " would surface.

[16] The failure to identify the presence of the initial vowel in such cases suggests that the same may have occurred elsewhere: hence, the absence of vowel initial verbs in the dictionary does not necessarily imply they were not encountered in the corpus; it would rather point to a failure of the parser to tear them out.

[17] Same situation in a Zimbabwean isiNdebele dictionary (Hadebe 2001)