New uses for old books: Description of digitised corpora-based on the Setswana language collection in the WITS Cullen Africana Collection

Rahlao, Malebogo
The University of the Witwatersrand, Johannesburg
thabonglebo@gmail.com

Lewin, Nina
The University of the Witwatersrand, Johannesburg
Nina.lewin@wits.ac.za

Surtee, Taariq
The University of the Witwatersrand, Johannesburg
Taariq.surtee@wits.ac.za

Abstract

This paper described a Corpus of 104 books. The books were catalogued into a standard library and archival metadata: Dublin core. A subset was digitised and cleaned. The books were then divided into five subsets and compared against each other and the entire Corpus. We speculated that the Corpus as a whole could be used as a general language register. Some examples are also given of the characteristics of the genre subsets. The paper aims to introduce the Corpus to natural language processing (NLP) researchers and offer it for further research.

1 Introduction

> *"The inability to communicate with human beings in their own language may be one of the most significant barriers to adopting information and communication technology in the third world and bridging the digital divide" (Getao & Evans, 2000: 128).*

At the same time, NLP is one of the most complex computational problems that have faced computer scientists. The calls for actions to develop resources for these languages, preferably in a coordinated and systematic manner, has been responded to in this paper. NLP researchers in English have had the advantage of many Corpora for research and testing. In this paper, we present the collection of the African language: Setswana sub-collection, for African Language NLP researchers in the 2nd Workshop on Resources for African Indigenous Languages.

1.1 Background

This project has a long history, from preserving books that would have been thrown out as outdated to creating a Corpus. We see it as unlocking the resources that many generations of archivists, librarians, researchers, authors, and translators have preserved in the belief that the books had value. Although, the individuals involved could not have imagined the computational uses in 1980. The William Cullen library: Africana Collection holds a subcollection of rare African language books dating from the 1800s in a locked area without researcher access. These books were originally the private collections of lecturers in the department. They were added to and carefully chosen by the African languages department as an internal library which was then given the Africana Library)[1]. These books provide a history of the development of various languages and varied orthography.

2 Collection Preparations

The first step was to catalogue the collection. There was an existing catalogue, but it was on physical cards and, we discovered, incomplete. We removed the books from the shelves and did a basic catalogue. We focused on creating a complete catalogue of books in Southern African languages. There is still an extensive collection in a variety of other African languages awaiting further research. The complete collection in Southern African languages is described elsewhere. We then compiled a metadata in Dublin core (and included some archival elements) of the Setswana books containing language varieties from Botswana and South Africa. The original catalogue designated these books as common readers. This paper describes the sub-collection of 104 Setswana books. These include Bibles, New and Old testaments, hymn books, prayer books, children's stories, grammar books, school books and novels.

[1] Margaret Atsango private correspondence 29-30/09/2021

3 Digitisation Methodology

The Avision book edge scanner in the Wits University Library Digitisation Centre converted all the materials into tiffs at high resolution. We used the international standards techniques and workflows adapted by the centre for the scanning of local content. After that, all the collected tiffs were cropped into Microsoft office 2010 documents. All scanned and cropped tiffs were then transferred into Abbyy FineReader 12 software. This software managed to split combined tiffs, skewed tiffs were deskewed, blank pages were removed, the pre-processing was applied, and lastly, they were saved in PDF format. The next step was to use the language recognition embedded in Abbyy FineReader 12 software to all the unrecognised PDFs by applying optical character recognition (OCR). This tool allowed the computer to recognise Setswana texts, moving from physical documents to text interpreted as data. This proves that the ground-breaking work of Otlogetswe in Setswana over many years (Otlogetswe, 2020, 2016, 2015, 2013, 2011a, 2011b, 2010, 2009a, 2009b, 2008; Otlogetswe and Chebanne, 2018; Otlogetswe and Ramaeba, 2014) which powered these recognition features is substantially effective. The recognition was challenging because the text was, like all older print, difficult, containing nonstandard fonts. Following the OCR process, we edited texts from the scanned documents as the OCR was approximately 80% effective in Setswana, with some unrecognised words. Such words were turned into different signs or symbols. Certain letters were misrecognised as other letters. We then cleaned the data by eliminating those signs and letters, replacing them with the correct Setswana letters manually. Pictures, text lines, and artefacts from the scanning process also had to be eliminated. The final product was then put into standard word documents.

4 Description of Corpora

The Corpus was created in Voyant (Sinclair & Rockwell, 2016). Basic descriptive tables as well as a Cirrus word cloud (see the appendices) were created. We created trends, collocations and correlation tables of the entire Corpus. To explore the linguistic varieties of the Corpus, we create five subsets: Bibles, Fairy tales, Grammar, Novels, and Poetry based on Genre.

The entire Corpus has 367149 total words and 23686 tokens. The appendices show a (10-line) sample of each.

Table 1: Example of Corpus Frequencies

Title	Words	Types	Ratio	Words/ Sentence
Bibles	57965	5164	0.089	17.984
Fairy tales	18980	2685	0.141	16.561
Grammar	9218	1774	0.192	10.030
Novels	34207	5799	0.169	19.435
Poetry	11550	3289	0.284	24.732

The average words per sentence in these subsets showcase Poetry as the highest (24.7) followed by Novels (19.4), Bibles (18.0), and Fairy tales (16.6), with Grammar (10.0) as the lowest. There is a case to made that the Corpus average of 19.3 could potentially be used as a language average, but this would require further research.

We also looked at the major terms

Table 2: Example of major terms

Term	RawFrequency	RelativeFrequency
batho	1305	3554.4153
kgosi	1169	3183.9934
modimo	967	2633.808
motho	764	2080.899
dira	690	1879.346
ja	618	1683.2402
bana	602	1639.6613
monna	579	1577.0165
utlwa	567	1544.3322
letsatsi	558	1519.8188
mosadi	502	1367.2922
bua	499	1359.1212
feta	475	1293.7527
tsaya	450	1225.6604
tloga	446	1214.7657
tsamaya	435	1184.805
ngwana	414	1127.6075
dikgomo	394	1073.1338
morafe	367	999.59424
motse	361	983.2521
lefatshe	354	964.1862

banna	315	857.9623
metsi	308	838.8965
tau	299	814.38324
baiseraela	276	751.7384
mokgwa	276	751.7384
lencwe	267	727.2252
mosimane	251	683.6461
nako	239	650.96185
ntlo	238	648.23816

4.1 Some features of the Corpus

The Corpus is able to support multiple types of research. The following were some initial features of interest that we found on an initial examination. The Corpus is still being annotated.

Orthography

This subset gives us information on how speech sounds form patterns. We were able to track changes because of the date range that our corpora spans. All our Bibles were written in an older orthography of the Serolong & Setlhaping dialect on a phonological level.. Our data also shows that most of our Bibles from the 1800s were written in the Serolong & Setlhaping dialect, known as the early orthographies of Setswana. Thus, early orthographies were based in Serolong & Setlhaping as the missionaries first created stellements among the Barolong & Batlhaping in Kudumane in 1821. In the older orthography, the letter [y] was used with the same pronunciation as the letter [j] used in the current orthography. For example, the words Yosefe (Joseph,), Yehofa (Jehova) and Yoshue (Joshua) are currently written as Josefa Jehofa and Jošua.

In the Serolong & Setlhaping dialect, people pronounce the phonemes / f / as [h] and / tsh / as [tšh]. For example, they pronounce the words 'world' and ‘resemblance’ as lehatshe and tšhobotsi, instead of lefatshe and tshobotsi.

Fairy Tales

Fairy tales are often intended for children, features fanciful and wondrous characters such as elves, goblins, wizards, and fairies. The term "fairy" tale refers more to the fantastic and magical setting or magical influences within a story rather than the presence of the character of a fairy within that story. The language register is simple because the writer speaks to children warning them in a casual and relaxed tone about a cannibal that eats children, for example. The vocabulary of the story is also easy for children to understand. Characters in fairy tales may be fairy folk or even talking animals, believable characters that children will care about such as a good-hearted hero, a scheming villain dimo (cannibal), in our example above), a wise helper, as so forth.

The word dimo (cannibal), is one of the highest frequencies (88) in the genre. Cannibals are often used in Setswana children's stories to scare them and warn them about the scheming villain that eats children. Many animals appear often in the genre. Setswana children's stories frequently consist of speaking animals that live with people. In most cases, an animal like nonyane (bird) is there to protect children from danger by keeping them under their wings while flying. Other examples of talking animals are phuduhudu (deer), phokojwe (jackal), and phiri (wolf). There are also human protagonists or participants in children's stories like Ntswakae, Ntitiagatsana, and Tsetsenyane. The word ja (eat) appears 43 times in the genre as children are often eaten in these stories.

It is clear that the register of the Corpus is mostly formal. In particular, the register used in the grammatical genre. It is academic in the sense that the grammatical textbook shows the writing style of the Setswana language.

We were keen to work on this Corpus especially because many other language collections depend on written speech such as newspapers, radio, or tweets that are relatively informal in register.

It would be interesting to compare this to (Marivate et al., 2020) newspaper Corpus.

5 Conclusion

This collection sat, from approximately the mid-1970s, on dusty shelves in a closed room. It was not used because the library did not have the language skills to catalogue it, and it was a fragile collection. This collection was nevertheless chosen by linguists because of its high value across a range of African languages. The language resources often available to NLP researchers are newspapers and social media forms that contain informal speech. This collection has well-described genres of formal writing. It is often asserted that there are insufficient African language texts available for research. However, as this work shows, the resources do exist but require work to convert into digital format.

The tables attached are samples to encourage an NLP researcher to explore the Corpus. We believe that our data (comparisons, colocations, and trends) can serve as a resource to pick tones. There is further work to be done because many of the books of the Corpus are translations. This means that the collection can also, with some work, provide line by line translations for test datasets.

Acknowledgements

Katlego Khoetha

eResearch Committee member

The University of the Witwatersrand, Johannesburg

Katlego.Khoetha@wits.ac.za

Katlego helped troubleshoot the original translation and assisted from a computing perspective.

Stéfan Sinclair

We would like to recognise the memory of Stéfan Sinclair, McGill University, who imagined, designed, and programmed much of Voyant Tools until 2020, when he passed away. See obituary.

Peter Duncan

Special Collections Librarian

William Cullen Library (retired)

The University of the Witwatersrand, Johannesburg

Peter curated and advocated for the collection to be catalogued and used throughout his career. The collection came to the library from the private collections of lecturers and departmental collections of the African Languages at WITS.

References and Bibliography

Aitchison, J., 1992. Teach yourself linguistics. Hodder & Stoughton., London.

Alexander, N., 1999. English Unassailable but Unattainable: The Dilemma of Language Policy in South African Education. Individual papers available online at http://www.

Atkins, S., 1991. Corpus design criteria. ICAME Journal 1–31.

Berg, A.S., 2018. Computational syntactic analysis of Setswana (Thesis). North-West University (South Africa), Potchefstroom Campus.

Bock, Z., Mheta, G., 2014. Language, society and communication an introduction.

Brits, K., Pretorius, R., Van Huyssteen, G.B., 2005. Automatic lemmatization in Setswana: towards a prototype. South African Journal of African Languages 25, 37–47.

Butt, M., King, T.H., 2003. Grammar writing, testing, and evaluation, in: Farghaly, A. (Ed.), Handbook for Language Engineers. CSLI, Stanford, California., pp. 129–180.

Evert, S., 2005. The statistics of word cooccurrences : word pairs and collocations (PhD Thesis). Universität Stuttgart, Stuttgart.

Getao, K.W., Miriti, E.K., 2006. Computational Modelling in Bantu Language.

Kubler, S., 2004. Memory-based parsing. J. Benjamins, Amsterdam.

Linguist, n.d. Setswana Scrabble | Sunday Standard [WWW Document]. URL http://www.sundaystandard.info/setswana-scrabble (accessed 7.29.19).

Louwrens, L.J., 1991. Aspects of Northern Sotho grammar. VIA Afrika, Pretoria.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., Modupe, A., 2020a. Low resource language dataset creation, curation and classification: Setswana and Sepedi -- Extended Abstract. arXiv e-prints arXiv:2004.13842.

Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., Modupe, A., 2020b. Low resource language dataset creation, curation and classification: Setswana and Sepedi -- Extended Abstract. arXiv:2004.13842 [cs].

Masoke-Kadenge, E., Kadenge, M., 2013. 'Declaration without implementation': An investigation into the progress made and challenges faced in implementing the Wits language policy. Language Matters 44, 33–50. https://doi.org/10.1080/10228195.2013.837949

Mogapi, K., 1998. The teaching of Setswana. Gaborone.

Molosiwa, A., Ratsoma, N., Tsonope, J., 1996. A comprehensive report on the use of Setswana at all levels of Botswana's education system, in: Cross-Border Languages. Reports and Studies. Regional Workshop on Cross-Border Languages, NIED, Okahandja, Namibia. pp. 99–134.

Nfila, B.I., 2006. Standard in Setswana in Botswana (Dissertation). University of Pretoria.

Nordlinger, R., Bresnan, J., 2011. Lexical-Functional Grammar: Interactions Between Morphology and Syntax 112–140.

Otlogetswe, T., 2020. Beef cuts amongst the Bangwaketse: the case of motlhakanelwa. Anthropology Southern Africa 43, 233–245. https://doi.org/10.1080/23323256.2020.1836667

Otlogetswe, T., 2016. The design of Setswana Scrabble. South African Journal of African Languages 36, 153-161–161.

Otlogetswe, T., 2015. Treatment of Spelling Variants in Setswana Monolingual Dictionaries. LEXI 25. https://doi.org/10.5788/25-1-1299

Otlogetswe, T., 2013. Introducing Tlhalosi ya Medi ya Setswana: The Design and Compilation of a Monolingual Setswana Dictionary. Lex 23. https://doi.org/10.5788/23-1-1228

Otlogetswe, T., 2011a. Challenges to Issues of Balance and Representativeness in African Lexicography. Lex 16. https://doi.org/10.5788/16-0-653

Otlogetswe, T., 2011b. Populating Sub-entries in Dictionaries with Multi-word Unitsfrom Concordance Lines. Lex 19. https://doi.org/10.5788/19-0-449

Otlogetswe, T., 2010. Challenges to issues of balance and representativeness in African lexicography. lex 16. https://doi.org/10.4314/lex.v16i1.51493

Otlogetswe, T., 2009a. Setswana Sports Terms: A Genre Analysis. Marang: Journal of Language and Literature 19. https://doi.org/10.4314/marang.v19i1.42819

Otlogetswe, T., 2009b. Populating sub-entries in dictionaries with multi-word units from concordance lines. lex 19. https://doi.org/10.4314/lex.v19i1.49140

Otlogetswe, T., 2008. Corpus design for Setswana lexicography (Thesis). University of Pretoria.

Otlogetswe, T., Chebanne, A., 2018. Setswana, in: Kamusella, T., Ndhlovu, F. (Eds.), The Social and Political History of Southern Africa's Languages. Palgrave Macmillan UK, London, pp. 187–221. https://doi.org/10.1057/978-1-137-01593-8_12

Otlogetswe, T., Ramaeba, G., 2014. Developing a Campus Slang Dictionary for the University of Botswana. Lex 24. https://doi.org/10.5788/24-1-1267

Pollard, C., Sag, I.A., 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press.

Pravec, N.A., 2002. Survey of learner corpora. ICAME journal 26, 8–14.

Pretorius, R., 2014. The sequence and productivity of Setswana verbal suffixes. Stellenbosch Papers in Linguistics Plus 44, 49–70. https://doi.org/10.5842/44-0-644

Pretorius, R.S., Posthumus, L.C., Kr??ger, C.J.H., Potchefstroom University for Christian Higher Education, 1997. Auxiliary verbs as a subcategory of the verb in Tswana.

Pretorius, R.S., Potchefstroom University for Christian Higher Education, 1997. Auxiliary verbs as a subcategory of the verb in Tswana.

Publications - Namibia Statistics Agency [WWW Document], n.d. URL https://nsa.org.na/page/publications (accessed 11.9.21a).

Publications - Namibia Statistics Agency [WWW Document], n.d. URL https://nsa.org.na/page/publications/ (accessed 11.9.21b).

Speech and Language Processing, 2nd Edition [WWW Document], n.d. URL https://www.pearson.com/content/one-dot-com/one-dot-com/us/en/higher-education/program.html (accessed 11.9.21).

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG, 2003. LANGUAGE POLICY.

Viljoen, B., Pretorius, L., Berg, A., Pretorius, R., 2008. Towards a computational morphological analysis of Setswana compounds. Literator : Journal of Literary Criticism, Comparative Linguistics and Literary Studies 29, 1–20.

Appendices – These are sample tables meant to introduce the types of data created from the Corpus.

Appendix 1 - Trends

Doc	Term	Raw	Relative	Z-Score	Z-Score	TF-		Distributions	
Index		Frequency	Frequency		Ratio	IDF			
0	ba	3647	62917.277	33.157654	-1017.6505	0.0	0.003450358	0.005520573	0.0055723283
4	ba	16906	45898.81	55.502193	-1703.4329	0.0	0.0065837344	0.005714953	0.00538373
2	ba	1471	43002.89	25.90773	-795.1412	0.0	0.00324495	0.0023679363	0.0019294296
0	le	2391	41249.03	21.703148	-666.09717	0.0	0.0059346156	0.0043819547	0.0039506597
1	go	378	41006.727	14.760752	-453.0262	0.0	0.004664786	0.0030375354	0.0042308527
2	go	1393	40722.66	24.528439	-752.80896	0.0	0.0038880932	0.004063496	0.003215716
4	le	14310	38850.816	46.971714	-1441.6216	0.0	0.0044199256	0.003885082	0.0037954887
5	ba	730	38461.54	13.78445	-423.06226	0.0	0.004056902	0.0072181243	0.0052687037
4	go	14079	38223.668	46.212646	-1418.3248	0.0	0.0031248983	0.0028588339	0.003700466
1	ba	344	37318.29	13.414559	-411.70987	0.0	0.004664786	0.0027120851	0.0018442179

Appendix 1 is a sample of Bible trends. There are ten distribution points. Three are presented here.

Appendix 2 – Collocated (example from Novels)

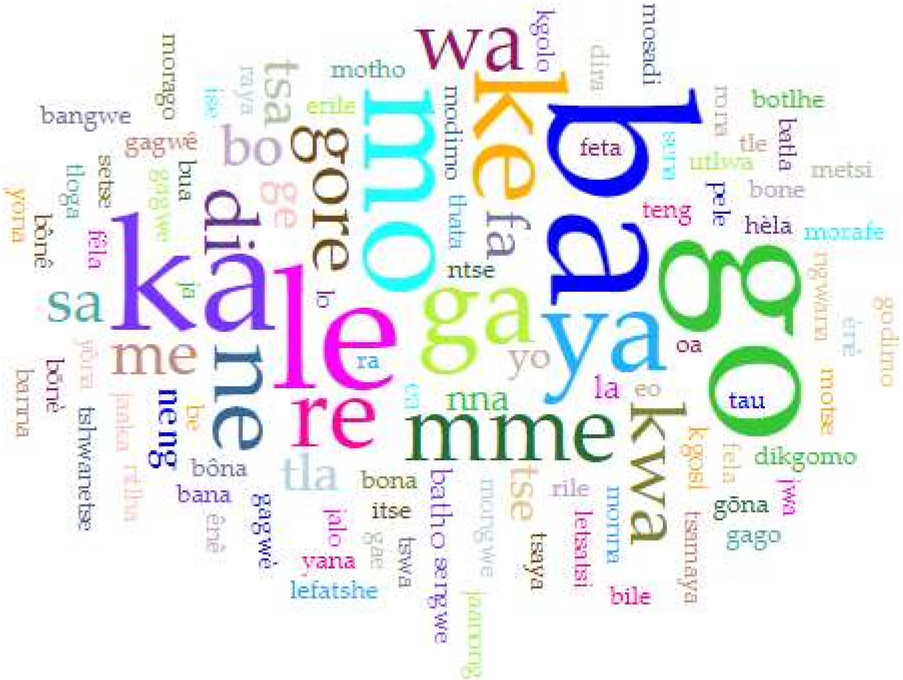
Term	Frequency	Context	Contextual Frequency
ba	1471	ba	1790
go	1393	go	736
le	1262	le	557
le	1262	go	502
go	1393	ba	497
ba	1471	go	486
ba	1471	le	480
go	1393	le	465
le	1262	ba	457
ke	802	ke	373

Appendix 3 – Correlation (example from Poetry)

Term 1	Term 2	Correlation
boÃªla	ngwale	1.0
matsodi	matsoke	1.0
diritibatsi	puiso	1.0
matute	morara	1.0
modikwadikwane	puiso	1.0
fofa	sefofu	1.0
ikaruse	sefofu	1.0
kitso	puiso	1.0
leo	matute	1.0
matute	semane	1.0

Appendix 4 – Cirrus word cloud – including the 105 most frequest words

Table 2 was derived from this list. We found the most frequent keywords.


wa
kgolo
mosadi
dira
motho
morago
raya
tsa
ise
erile
botlhe
bangwe
bo
gore
mo
modimo
ke
feta
ba
go
rona
tle
batla
utlwa
metsi
gagwe
bua
setse
tloga
yona
bônê
fêla
di
ge
thata
fa
pele
teng
bone
hèla
morafe
ngwana
ntse
ka
le
ga
ya
sa
a
lo
ênê
godimo
oa
me
ra
yo
la
eo
tau
motse
ne
re
mme
kwa
nna
ka
be
neng
bana
bôna
tla
bona
itse
batho
sengwe
mongwe
tse
rile
letsatsi
moma
kgosi
fela
dikgomo
iwa
gôna
gago
tsamaya
tsaya
bile
jalo
yana
gae
tswa
lefatshe
gagwe
ênê
jaanong
tshwanetse
jaaka
ntlha
bana
jôna