Journal of Open
Humanities Data

A Data Set of Final Year High School Examination Texts of South African Home and First Additional Language Subjects

DATA PAPER

JOHANNES SIBEKO
MENNO VAN ZAANEN
*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

This article describes a data set of reading comprehension and summary writing texts that were used in final-year high school examinations in South Africa between 2008 and 2020. It contains texts for eleven official South African languages. PDF versions of the texts stem from South Africa's Department of Basic Education's online public access repository. Plain text is extracted from the PDFs and the texts are tokenized. The data set contains 429 full-text files with 929 manually extracted comprehension and summary writing texts. The data is useful for studies investigating, e.g., linguistic properties, text readability, text properties, and linguistic complexity in any of the eleven languages. Furthermore, both intra-language and inter-language comparisons or investigations can be made.

CORRESPONDING AUTHOR:
Johannes Sibeko
Linguistics and Applied Linguistics, Nelson Mandela University, Gqeberha, South Africa
Johannes.Sibeko@mandela.ac.za

KEYWORDS:
linguistic corpus; indigenous languages; examination texts; reading comprehension; summary writing; final year high school

TO CITE THIS ARTICLE:
Sibeko, J., & van Zaanen, M. (2023). A Data Set of Final Year High School Examination Texts of South African Home and First Additional Language Subjects. *Journal of Open Humanities Data*, 9: 9, pp. 1–6. DOI: https://doi.org/10.5334/johd.108

1 OVERVIEW

REPOSITORY LOCATION

South African Centre for Digital Language Resources: https://repo.sadilar.org/; data set: https://hdl.handle.net/20.500.12185/568.

CONTEXT

School texts, e.g., reading comprehension (François & Fairon, 2012) or language instruction texts (Curto, Mamede, & Baptista, 2014, 2015; Forsyth, 2014), have been historically used in complexity studies. We leverage the reading comprehension and summary writing texts from examination question papers to overcome the limitations of reproduction of copyrighted textbook materials. We utilise texts from the home language (HL) and the first additional language (FAL) examinations. The home language subject is aimed at learners who start the first grade with competencies such as reading, writing, speaking, and listening in the language (Department of Basic Education, 2011b). The first additional language subject is tailored for learners who do not necessarily start the first grade with competencies or exposure to the language being taught (Department of Basic Education, 2011b). According to Makalela (2023), the objectives of the two subject levels are largely similar. However, the texts administered to learners in the HL subject are more linguistically complex (Sibeko, 2021) and are harder to read than those in the FAL classes, at least as far as English is concerned (Sibeko & van Zaanen, 2021). The current data set has already been used in the following articles:

- Sibeko, J. (2021). A comparative analysis of the linguistic complexity of grade 12 English Home Language and English First Additional Language examination papers. *Per Linguam: a Journal of Language Learning*, *37*(2), 50–64. DOI: https://doi.org/10.5785/37-2-976
- Sibeko, J., and van Zaanen, M. (2021). An analysis of readability metrics on English examination texts. *Journal of the Digital Humanities Association of Southern Africa*, *03*(1), 1–11. DOI: https://doi.org/10.55492/dhasa.v3i01.3864

2 METHOD

STEPS

The data collection process consisted of four steps. First, PDF files of the examination papers were downloaded from South Africa's Department of Basic Education's website.[1] As such, no student responses are available. These files (like all other files in the data set) are manually organized per language, per subject (either HL or FAL), and per examination opportunity. Language examinations are written in three sections, i.e., paper one for language, paper two for literature, and paper three for creative writing[2] (Department of Basic Education, 2011a, 2011b, 2011c). Second, plain text was extracted from the PDF files using pdftotext (version 22.02.0), which is language-independent, on an Ubuntu Linux platform. Third, the plain texts were tokenized (and sentencized) using Ucto (version 0.21.1) to identify the individual words and sentences in the texts. These are both open-source tools. Fourth, the reading comprehension and summarization texts were manually extracted from the tokenized plain text files. Note that some examination papers contain more than one reading comprehension text. The names of all text files contain relevant metadata (language, subject, year, month, and file type).

Table 1 provides an overview of the distribution of the files in the data set. Table 2 provides an overview of the token and type (i.e., unique tokens) counts of the full examination texts, whereas Table 3 provides the same information for the extracted reading comprehension and summarization texts.

The data set contains 429 full examination text files. Of these, 223 are HL texts that have 689,730 tokens and 88,009 types, whereas the 206 FAL text documents contain 624,821 tokens with 73,451 types. In addition to the full examination texts, the reading comprehension and summary writing text part of the examinations are extracted manually, resulting in 929 texts (481 for HL and 448 FAL texts) with 472,430 tokens and 87,779 types. The extracted HL

[1] https://www.education.gov.za [Last accessed: 27 June 2023].

[2] We only use text from paper one as paper two contains copyrighted material and paper three only provides a prompt for the creative writing assignment.

Table 1 Distribution of texts per language and subject level for both examination texts and extracted reading comprehension and summarization texts.

	# OF EXAMINATION TEXTS			# OF EXTRACTED TEXTS		
	HL	FAL	TOTAL	HL	FAL	TOTAL
Afrikaans	21	22	43	53	58	111
English	24	25	49	56	57	113
IsiNdebele	20	16	36	43	34	77
IsiXhosa	19	21	40	39	42	81
IsiZulu	18	18	36	36	39	75
Sepedi	22	20	42	48	42	90
Sesotho	22	19	41	49	39	88
Setswana	17	14	31	34	29	63
Siswati	21	18	39	43	38	81
Tshivenḓa	19	17	36	39	37	76
Xitsonga	20	16	36	41	33	74

Table 2 Token and type count per language and subject level for the full examination texts.

	# OF TOKENS IN EXAMINATION TEXTS			# OF TYPES IN EXAMINATION TEXTS		
	HL	FAL	TOTAL	HL	FAL	TOTAL
Afrikaans	77,787	90,731	168,518	8,943	6,946	12,829
English	80,252	86,113	166,365	8,497	7,489	12,325
IsiNdebele	48,931	37,519	86,450	12,430	9,719	18,903
IsiXhosa	50,480	53,518	103,998	13,529	13,488	23,136
IsiZulu	43,456	44,082	87,538	11,738	11,076	19,726
Sepedi	66,846	56,594	123,440	5,709	5,253	8,374
Sesotho	80,592	66,934	147,526	6,900	5,811	9,738
Setswana	52,836	42,026	94,862	5,587	4,580	8,106
Siswati	54,597	43,845	98,442	14,608	10,792	21,868
Tshivenḓa	62,726	52,881	115,607	5,694	4,636	7,877
Xitsonga	71,227	50,579	121,806	5,831	4,446	7,933

Table 3 Token and type count per language and subject level for the extracted reading comprehension and summarization texts.

	# OF TOKENS IN EXAMINATION TEXTS			# OF TYPES IN EXAMINATION TEXTS		
	HL	FAL	TOTAL	HL	FAL	TOTAL
Afrikaans	29,298	24,804	54,102	5,761	3,955	8,019
English	33,625	27,131	60,756	6,064	4,900	8,599
IsiNdebele	18,346	12,637	30,983	7,890	6,146	12,268
IsiXhosa	19,980	18,601	38,581	8,920	8,611	15,347
IsiZulu	18,812	15,540	34,352	8,211	6,732	13,177
Sepedi	27,439	19,235	46,674	3,995	3,444	5,845
Sesotho	29,754	21,182	50,936	4,456	3,413	6,235
Setswana	21,303	14,269	35,572	3,683	2,741	5,289
Siswati	20,885	13,607	34,492	8,902	6,300	13,394
Tshivenḓa	25,473	18,693	44,166	3,889	2,900	5,273
Xitsonga	25,932	17,778	43,710	3,899	2,904	5,405

texts consist of 269,881 tokens and 59,007 types, whereas the extracted FAL texts consist of 202,549 tokens and 46,356 types.

The data set is combined into one ZIP file. The files in the data set are first divided into their file type (directories called pdf for PDF files, txt for the extracted UTF-8 Unicode files, tok for the corresponding tokenized files, and ext for the manually extracted reading comprehension and summarization texts). Next, the files are divided into directories corresponding to their languages. Within these directories, the files are divided into directories describing the two subjects, namely, HL and FAL. Next, the files are divided into three examination months, namely, February or March, May or June, and November, respectively in Feb, May, and Nov directories. The files have a consistent naming scheme: *lang_subj_month_year.ext* with *lang* the name of the language, *subj* the name of the subject level, *month* either Feb-March, May-June, or Nov, depending on the months the examinations were written, *year* ranging from 2008–2020. *ext* represents the type of the filename, either txt for text files, or pdf for PDF files. For the extracted reading comprehension and summarization files (found in the ext directory), before the extension (.*ext*), _*type* is present. This *type* can take the values RC1, RC2 for the first and second reading comprehension texts respectively, or SUM for the summarization texts. For instance, a filename IsiZulu_FAL_Nov_2009_SUM.txt indicates a summary text from an IsiZulu FAL examination written in November in the year 2009.

SAMPLING STRATEGY

All available examination texts have been downloaded from South Africa’s Department of Basic Education’s website. However, as can be seen in Table 1, for some languages certain examination texts have not been made available.

One full examination text consists of reading comprehension texts in the first section, summary writing texts in the second section, and visual texts and language convention texts in the third section. We excluded the third section as it regularly contained graphics, such as cartoons or advertisements, and often it contained deliberate errors.

QUALITY CONTROL

The authors manually checked the contents of the texts to ensure all sections found in the PDF documents are also found in the plain text variants. Additionally, the texts were checked for consistent use of text encodings, in particular, related to diacritics for Tshiveṋḓa and Afrikaans.

3 DATA SET DESCRIPTION

OBJECT NAME

Final year high school examination texts of South African home language and first additional language subjects.

FORMAT NAMES AND VERSIONS

PDF, UTF-8 encoded text files.

CREATION DATES

Start date: 2021–02–01; End date: 2022–10–15.

DATA SET CREATORS

Johannes Sibeko and Menno van Zaanen.

LANGUAGE

The data set contains texts in Afrikaans, English, isiNdebele, isiXhosa, isiZulu, Sepedi, Sesotho, Setswana, Siswati, Tshiveṋḓa, and Xitsonga. Metadata is provided in English.

LICENSE

Creative Commons License Attribution-ShareAlike 4.0 International.

REPOSITORY NAME

South African Centre for Digital Language Resources.

PUBLICATION DATE

2022–11–23.

4 REUSE POTENTIAL

Research in linguistic and text complexity (related to readability of texts) has been on-going for over a century (Collins-Thompson, 2014; De Clercq & Hoste, 2016). However, such research on South African languages has lagged behind, resulting in limited resources for analysing readability and complexity of texts in the indigenous languages (Sibeko & De Clercq, 2023).

The texts in the data set allow for several linguistic comparisons: chronologically or diachronically, between languages (i.e., cross-lingual comparison), between subjects (HL versus FAL), between types of texts (summary versus reading comprehension), and between different examination dates (February versus May versus November). Further annotation of the texts (e.g., on part-of-speech, partial parsing, named entities, etc.) allows for investigations into these textual properties. As the data set contains texts of languages of both disjunctive and conjunctive orthographies, investigations into the influence of orthography can be performed. For instance, the orthography of isiZulu has been proven to affect reading ability (Land, 2015).

More content-oriented research can consider the different genres used. For instance, around 2008 to 2011, literary texts were used for reading comprehension (taken from books that are not part of the official curriculum in the language), but after 2012, these texts focused more on newspaper and magazine news articles. This allows for research into the influence of the different genres (Solovyev, Ivanov, & Solnyshkina, 2018).

The data set is, to our knowledge, the first corpus in the South African educational context. The data allows for research in the realm of education. As examples, we mention, investigations into the themes of the texts used for the different languages and subjects, investigations into overall learner achievements between similar languages (e.g., those in the Nguni or the Sotho group), and investigations into learners' reading abilities as the current Progress in International Reading Literacy Study (PIRLS) indicate depreciating reading abilities through the years (Mullis, von Davier, Foy, Reynolds, & Wry, 2023).

ACKNOWLEDGEMENTS

This article forms part of Johannes Sibeko's doctoral research project conducted at North-West University. The research is being carried out under the guidance and supervision of Menno van Zaanen

FUNDING INFORMATION

This research was funded in part by the National Research Foundation (NRF) of South Africa (Grant number: 128875).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Johannes Sibeko – Conceptualization, Data curation, Funding acquisition, Methodology, Validation, Writing – original draft, Writing – review & editing, Writing – final draft.

Menno van Zaanen – Conceptualization, Data curation, Methodology, Formal Analysis, Validation, Supervision, Writing – review & editing, Writing – final draft.

AUTHOR AFFILIATIONS

Johannes Sibeko orcid.org/0000-0003-3586-7491
Linguistics and Applied Linguistics, Nelson Mandela University, Gqeberha, South Africa

Menno van Zaanen orcid.org/0000-0003-1841-2444
South African Centre for Digital Language Resources, North-West University, Potchefstroom, South Africa

REFERENCES

Collins-Thompson, K. (2014). Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics*, *165*(2), 97–135. DOI: https://doi.org/10.1075/itl.165.2.01col

Curto, P., Mamede, N., & **Baptista, J.** (2014). Automatic readability classifier for european portuguese. *System*, *5*(1), 309–324.

Curto, P., Mamede, N., & **Baptista, J.** (2015). Automatic text difficulty classifier: Assisting the selection of adequate reading materials for european portuguese teaching. In M. Helfert, M. T. Restivo, S. Zvacek, & J. Uhomoibhi (Eds.), *Proceedings of the 7th international conference on computer supported education (csedu-2015), 1*, 36–44. Science and Technology Publications. DOI: https://doi.org/10.5220/0005428300360044

De Clercq, O., & **Hoste, V.** (2016). All mixed up? finding the optimal feature set for general readability prediction and its application to English and dutch. *Computational Linguistics*, *42*(3), 457–490. DOI: https://doi.org/10.1162/COLI_a_00255

Department of Basic Education. (2011a). *Curriculum and assessment policy statement: English first additional language grades 10–12*. Pretoria: Government Printing Works.

Department of Basic Education. (2011b). *Curriculum and assessment policy statement: English home language grades 10–12*. Pretoria: Government Printing Works.

Department of Basic Education. (2011c). *Curriculum and assessment policy statement: English second additional language grades 10–12*. Pretoria: Government Printing Works.

Forsyth, J. N. (2014). *Automatic readability detection for modern Standard Arabic* (Thesis). Bringham Young University.

François, T., & **Fairon, C.** (2012). An «AI readability» formula for French as a foreign language. In *Proceedings of the 2012 joint conference on empirical methods in Natural Language Processing and computational natural language learning* (pp. 466–477). Association for Computational Linguistics.

Land, S. (2015). Reading and the orthography of isiZulu. *South African Journal of African Languages*, *35*(2), 163–175. DOI: https://doi.org/10.1080/02572117.2015.1113000

Makalela, L. (2023). Merging English Home Language and First Additional Language curricula: Implications for future quality assurance practices. *Southern African Linguistics and Applied Language Studies*, *41*(1), 76–87. DOI: https://doi.org/10.2989/16073614.2023.2185984

Mullis, I. V. S., von Davier, M., Foy, B., P Fishbein, Reynolds, K. A., & **Wry, E.** (2023). *Pirls 2021: International results in reading*. Boston College: TIMSS & PIRLS International Study Center. DOI: https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Sibeko, J. (2021). A comparative analysis of the linguistic complexity of grade 12 English Home Language and English First Additional Language examination papers. *Per Linguam*, *37*(2), 50–64. DOI: https://doi.org/10.5785/37-2-976

Sibeko, J., & **De Clercq, O.** (2023, May 2–6). A corpus-based list of frequently used words in Sesotho. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages*. Dubrovnik, Croatia: Association for Computational Linguistics.

Sibeko, J., & **van Zaanen, M.** (2021). An analysis of readability metrics on English exam texts. *Journal of the Digital Humanities Association of Southern Africa*, *03*(1), 1–11. DOI: https://doi.org/10.55492/dhasa.v3i01.3864

Solovyev, V., Ivanov, V., & **Solnyshkina, M. I.** (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. *Journal of intelligent and fuzzy systems*, *34*(5), 3049–3058. DOI: https://doi.org/10.3233/JIFS-169489

TO CITE THIS ARTICLE:
Sibeko, J., & van Zaanen, M. (2023). A Data Set of Final Year High School Examination Texts of South African Home and First Additional Language Subjects. *Journal of Open Humanities Data*, 9: 9, pp. 1–6. DOI: https://doi.org/10.5334/johd.108

Submitted: 27 April 2023
Accepted: 09 June 2023
Published: 05 July 2023