Creating electronic resources for African languages: challenges and opportunities

E. Taljard, D.J. Prinsloo, M .Goosen
15 February 2023

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Humanities
Fakulteit Geesteswetenskappe

Make today matter

Introduction and aims (1/2)

- The need for electronic resources for (under-resourced) African languages
- Development of Human Language Technology (HLT)
- These technologies rely on large quantities of high-quality electronic data
- Digitisation understood as
  - the conversion of analogue text, audio and video data into digital form
  - the provision of born digital data

Introduction and aims (2/2)

- Information on UP digitisation node on:
  - tools, procedures, best practices and standards
  - digitisation of text, audio and audio-visual material for the African languages
- SADiLaR (https://www.sadilar.org/index.php/en/)

Text digitisation

- Selection of text to be digitised
- Hardware and software used for digitisation
- Experiment: comparing OCR software using Afrikaans, isiZulu, Sepedi and Tshivenḓa for illustrative purposes:
  - ABBYY FineReader 14
  - Omnipage Professional 18, and
  - CTexTools

Average accuracy rate of software

Percentage of scanning errors and overall accuracy rate

	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**

- OCR – conclusion: ABBYY FineReader 14 at 300 dots per inch (dpi)

Correction of defects

- Pre-processing,
- Split facing pages,
- Deskew images,
- Straighten text lines,
- Crop,
- Remove colour marks,
- Etc.

Text cleaning: web-sourced material 1/2

- Web-sourced material consists of web pages, i.e. documents marked up by HTML
- Web pages contain boilerplate texts:
  - navigational structures, e.g. menu’s;
  - headers, e.g. logos & breadcrumbs;
  - footers, e.g. copyright notices & dates; and
  - advertisements

Text cleaning: web-sourced material 2/2

- Cleaning of web-sourced texts needs to be maximally automatized – manual cleaning not feasible – neither timewise nor in terms of human resources
- Several solutions have been suggested; software not yet been evaluated for African languages
- Example: NCLEANER, tool for automatic boilerplate removal, uses character-level N-gram models as classifiers
- Limited applicability of cleaning tools for African languages

Text cleaning: text-based material 1/6

- Cleaning of (hard copy) text-based material requires less specialized computational knowledge than cleaning of web-sourced material
- Three major types of scanning errors:
  - Duplication
  - Orthographical, spelling & word division errors in original texts
  - Basic scanning errors

Text cleaning: text-based material 2/6

- Focus on correction of scanning errors
- Text cleaning strategies include:
  - manual correction;
  - spellchecker support;
  - automatic search & replace individual items;
  - automatic search and replace with basic macros;
  - detecting and cleaning duplications through concordance line repetitions; and
  - anonymizing texts containing sensitive information.

Text cleaning: text-based material 3/6

Detecting & cleaning duplications

1	Sekolo se sa Realeka ga se sekolo se	Sekolo se sa Realeka ga se sekolo se segolo k
2	botho, gape ba fiwa mešomo ye bothata. Sekolo se sa Realeka ga se sekolo se	Sekolo se sa Realeka ga se sekolo se segolo k
3	ba maatla ka kua. Sekolo ga se selo, gape sekolo ga se na mohola!	Go ya sekolong.txt
4	le bašemane ba bagolo ba maatla ka kua. Sekolo ga se selo, gape sekolo ga se na	Go ya sekolong.txt
5	ye bothata. Sekolo se sa Realeka ga se sekolo se segolo kudu. Go na le	Sekolo se sa Realeka ga se sekolo se segolo k
6	Sekolo se sa Realeka ga se sekolo se segolo kudu. Go na le	Sekolo se sa Realeka ga se sekolo se segolo k
7	. Ke nyaka go ya sekolong se segolo, sekolo sa Realeka Go na le bašemane ba	Go ya sekolong.txt
8	fela. Morutiši le bana ga ba kwane ka tša sekolo. Morutiši o re o rata ngwana yo	Sekolo se sa Realeka ga se sekolo se segolo k
9	fela. Morutiši le bana ga ba kwane ka tša sekolo. Morutiši o re o rata ngwana yo	Sekolo se sa Realeka ga se sekolo se segolo k

Text cleaning: text-based material 4/6

Anonymisation of texts

- Necessity of sharing data; limitations posed by data containing sensitive information
- Data anonymization = process of masking or removing sensitive data so that it is no longer personal
- Autshumato Text Anonymizer: “tool for the anonymisation of text corpora which entails the identification of entities that may convey confidential information and replacing those entities with randomly selected entities of the same type”

Text cleaning: text-based material 5/6

Anonymisation of texts cont.

- Two critical challenges for anonymization tools:
  - Anonymized text must remain readable, meaningful & useful for syntax & content-based analysis
  - Evaluation of performance of anonymizing tool is challenging
- Mixed results with the use of the Autshumato tool:

Text cleaning: text-based material 6/6

Anonymisation of texts cont.

INPUT:

Lehono ke Lamorena. Thato le Nnake ba ya kerekeng le batswadi ba bona. Bašemane ba apere diaparo tša bona tša Lamorena. Leina la moruti ke Thapelo Mojela. O ithutile boruti Yunibesithing ya Pretoria.

Mongwadi ke Elsabé Taljard, leina la Sepedi la gagwe ke Mphakiseng.

OUTPUT 1:

Lehono ke <DATE type="days">Lamorena</DATE>. <NE type="firstnames">Thato</NE> le Nnake ba ya kerekeng le batswadi ba bona. Bašemane ba apere diaparo tša bona tša <DATE type="days">Lamorena</DATE>. Leina la moruti ke <NE type="firstnames">Thapelo</NE> Mojela. O ithutile boruti Yunibesithing ya Pretoria.

Mongwadi ke Elsabé Taljard, leina la Sepedi la gagwe ke Mphakiseng.

OUTPUT 2:

Lehono ke Labone. Faricah le Nnake ba ya kerekeng le batswadi ba bona. Bašemane ba apere diaparo tša bona tša Mokibelo. Leina la moruti ke Minnelise Mojela. O ithutile boruti Yunibesithing ya Pretoria.

Mongwadi ke Elsabé Taljard, leina la Sepedi la gagwe ke Mphakiseng.

Granularity of cleanness

Different granularity levels of cleanness of corpora enable different application possibilities

Application of corpus	Status/quality/condition of corpus	Required correction methods
Frequency lists	Dirty corpus	No correction necessary, use as is
Authentic (corpus) examples	Dirty corpus	No correction necessary, use as is
Concordance lines (keyword-in-context)	Dirty corpus	No correction necessary, use as is
Text verification - part of speech (POS) matches	Clean corpus	Semi-automatic spelling checking or semi-automatic search and replace operations
Mark-up: e.g. POS, morphological analysis, lemmatisation	Clean corpus	Semi-automatic spelling checking or semi-automatic search and replace operations
Rare occurrences of words	Relatively clean corpus	Correction of typical OCR errors by semi-automatic search and
Spelling checkers, Grammar checkers, Text verification - exact matches	100% clean corpus	Correction of all OCR errors by proofreading of the text or at least semi-automatic spelling checking or semi-automatic search and replace operations

Digitisation of audio material / cassettes (1/2)

- Hardware and software used for digitising audio material
  - USB Cassette Capture (tape to MP3 converter)
  - Audacity (https://www.audacityteam.org/about/)
- Determining the quality of audio cassettes
- Quality affected by
  - incorrect storage
  - deterioration because of age and
  - physical damage to cassettes

Digitisation of audio material / cassettes (2/2)

- Default quality settings
  - 44100 Hertz (Hz) and 32-bit format in stereo
  - Digitised file format: .aup
  - Final format stored in Waveform Audio File Format (WAV file, .wav)

Digitisation of video material

- Old technology: technical challenges
- Hardware and software used for digitising video material
  - Video Home Systems (VHS) and Elgato video capture
- Very specific criteria for digitisation
  - resolution, bitrate, audio, colour mode, recording format, recording video type, On Screen Display (OSD) messages
  - digitised version stored in .mpg (MPEG2) format (codec: MPEG-2 video (mpgv) / codec: MPEG audio layer 2 (mpga))

Provision of metadata (1/3)

- 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”.
- Should be presented in an integrated form, together with the text file
- Facilitates the identification, management, access, use and preservation of a digital resource
- TEI (Text Encoding Initiative)

Provision of metadata (2/3)

- Provision of a standard bibliographic description
  - Texts: title, name of author(s), date of publication, ISBN, publisher, genre, description of genre, language (using ISO language codes), status of copyright, number of pages (PDF), tokens, media type, encoding, format / file extension and name of document
  - File names:
    - Creative text, e.g. novel, drama, poetry, etc.: *zul_Zibukhipha zibuthela_Shabangu_novel*
    - Newspaper / magazine: *sot_Boleng ba popeho tse fapaneng tsa mebu_Pula Imvula_201210na*

Provision of metadata (3/3)

- Video and audio material: title, presenter, date of publication, publisher, genre, description of genre, language (using ISO language codes), status of copyright, length, media type, encoding, format / file extension and name of document
- File names: *afr_AFR 102 verstegniek onderrigkasset 3 kant A_Marais_19990315*

Copyright considerations

- Copyright as salient aspect of text digitisation
- Digitisation is a form of (re)publishing
- Obtaining copyright clearance is essential
- Fair use
- No safe generic copyright rules for text scanning
  - obtain explicit permission of the copyright owner

Conclusion

- Digitisation entails much more than simple scanning
- Data stored in the correct format
- Follow the correct
  - Protocols
  - Procedures
  - Technical guidelines
- Preservation of material
- Quality-compromised exceptions

Acknowledgement

- This research is supported by the South African Centre for Digital Language Resources (SADiLaR)

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.
- Nicholson, D. 2010. Copyright and related matters. In Liebetrau (red.) 2010.
- Prinsloo, D.J., Taljard, E. and Goosen, M. 2022. Optical Character Recognition and text cleaning in the indigenous South African languages. SpilPlus Vol 64, 165 -187
- Senekal, 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.