

# African Language Information Retrieval



*Hussein Suleman*

*Digital Libraries Laboratory @ Centre for ICT4D  
Department of Computer Science, School of IT, UCT*

*March 2019*



# University of Cape Town





# Digital Libraries Lab @UCT

- Making information available to people
- Context-sensitive: low resource environment, different languages, different skills/culture, etc.
- 5 areas of interest:
  - Information Retrieval (IR) – HS
  - Digital Libraries (DL) - HS
  - Educational Technology – HS
  - Ontologies/Knowledge engineering – CMK
  - NLP - CMK
- Some production systems / products





# What is Information Retrieval (IR)?

- Basically, Search Engines
- Also:
  - Recommender systems
  - Classification
  - Sentiment/Topic analysis
  - Some data mining
  - Some databases





# Goal: IR and Human Development

## □ Human Dignity

- Promote the status of local languages.
- Create tools that support local languages.
- Increase presence of local languages.

## □ IR4D

- IR for employment, governance, health, etc.





## IR in Africa

- ❑ Little algorithmic support in IR/NLP.
- ❑ Very little and noisy data.
- ❑ <1000 Wikipedia documents.
- ❑ Unclear language boundaries.
- ❑ Society not search savvy.
- ❑ Multilingualism.
- ❑ Mixed queries/documents.
- ❑ Resource limitations.





## Arabic IR

- Feasibility study in 2006 into IR in isiXhosa, isiZulu, etc.
  - Very few text resources online.
  - Very little research in related areas.
  - Virtually no IR research.
- Arabic as African language!
  - Spoken in most of North Africa.
  - Dialect/usage not the same as Middle East.
  - Fashionable at the time.





# Mixed Language IR

*Mohammed Mustafa Ali, PhD*

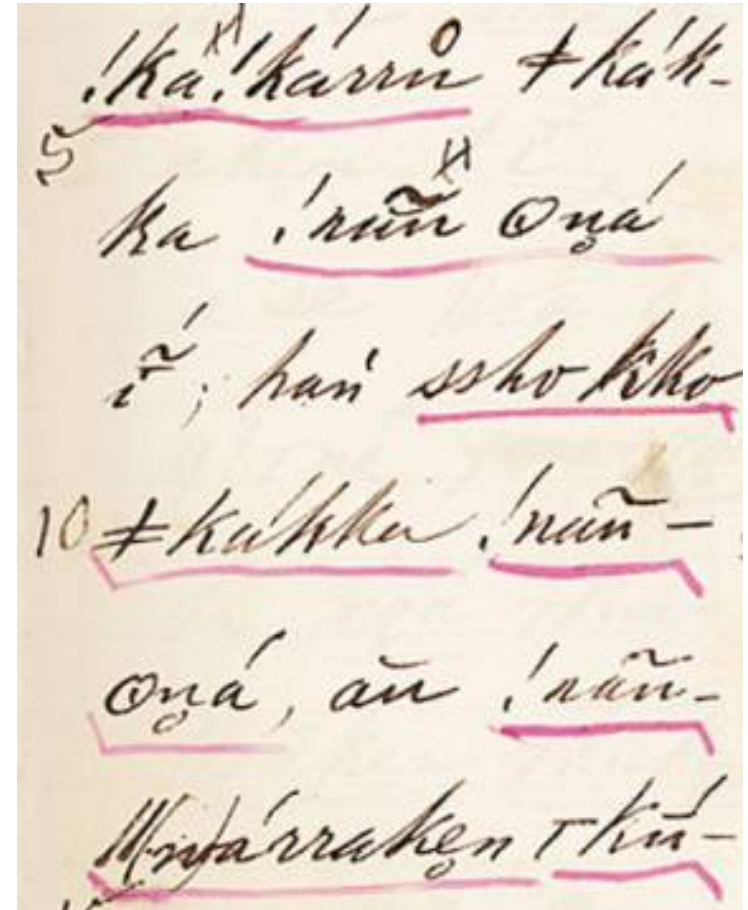
- Noted that Google is language unaware.
- Poor results for mixed queries – queries in multiple languages.
  - Dominant languages are dominant in results.
  - Mixed language use is very popular in Africa.
- Solution: Examine queries and rerank based on language-based collection weights.





## |Xam IR

- ❑ Extinct Khoisan language.
- ❑ Language used in documenting early South African history/culture (25000 pages of stories).
- ❑ No Unicode representation.





# Digital Bleek and Lloyd Collection



## THE DIGITAL BLEEK AND LLOYD

## HOME

This digital publication is part of a Llarec project to digitise, research and publish the Bleek and Lloyd Archive. The Digital Bleek and Lloyd includes scans of every page of the 110 Lucy Lloyd !xam notebooks, 17 Lloyd (mostly) !kun notebooks and 28 Wilhelm Bleek !xam notebooks. It also includes Jemima Bleek's solitary Korana and !kun notebook and four Lloyd Korana notebooks in the Maingard collection of the Library at the University of South Africa, as well as Dorothea Bleek's 32 notebooks. All the drawings and watercolours made by !han#kass'o, Dia!kwain, Tamme, Juma, !nanni and Da are also in the digital collection. The digital archive includes a 280 000-word searchable index, cross-referenced and including notes and summaries for each of the stories listed. Notes in italics are direct quotes from the reports of Bleek and Lloyd in which they detailed the progress of their research.

Llarec (the Lucy Lloyd Archive, Resource and Exhibition Centre) is part of the Centre for Curating the Archive, a University of Cape Town research centre directed by Pippa Skotnes and located at the Michaelis School of Fine Art. The initial "Digital Bleek and Lloyd" accompanied the publication "Claim to the Country: the Archive of Wilhelm Bleek and Lucy Lloyd" by Pippa Skotnes (2007), published by Jacana Media and Ohio University Press. Subsequently Jemima Bleek's and Dorothea Bleek's notebooks have been added, as well as the Digital Stow, featuring the rock art copies of George Stow. The search index and summaries have also been extended and currently the Bleek and Lloyd dictionaries are being digitised. Please refer to the CCA website at <http://www.cca.uct.ac.za> for updates.

The project has been made possible by funding provided by the Andrew W. Mellon Foundation and De Beers; and is the result of the cooperation of the four curating institutions: University of Cape Town, Unisa, Iziko South African Museum and The National Library of South Africa.

**These scans of the documents and images that comprise the Bleek and Lloyd archive may not be used or reproduced for any purpose without permission of the copyright holders.**



### Books

Cover to cover | Contributor | Story | Category | Keyword



### Drawings/Watercolours

Contributor | Category | Keyword



### Digital George Stow

Images



### Contributors

!xam | !kun | Researchers





# Bleek and Lloyd: Low Resource IR

- IR engine within the browser – no network needed.
- Only simple transcriptions supported.

The screenshot displays the 'THE DIGITAL BLEEK & LLOYD' website. On the left is a navigation menu with categories like 'HOME', 'Notebooks', 'Drawings/Watercolours', 'Digital George Stow', and 'Contributors'. The main content area shows a search results page with the query 'ixam'. The search results are listed as follows:

1. [Notebooks / Story Index / The Ixam and the Dutch stories/19/index.html](#)
2. [Notebooks / Story Index / Ixam-speaking people stories/55/index.html](#)
3. [Notebooks / Story Index / Difference between Ixam and European methods of articulation stories/889/index.html](#)
4. [Notebooks / Story Index / Bushman genealogies stories/30/index.html](#)
5. [Notebooks / Story Index / The personal histories of various people stories/583/index.html](#)
6. [Notebooks / Story Index / A fragment about the animal clicks, and ways of speaking Ixam stories/326/index.html](#)
7. [Notebooks / Story Index / Words and sentences: Ixam names of animals stories/790/index.html](#)
8. [Notebooks / Story Index / Ixam names stories/715/index.html](#)





# Bleek and Lloyd: Dictionary

*Lebogang Molwantoa, Sanvir Manilal, Kyle Williams, BSc(Hons)*

- Visual dictionary – pictures of words.
- Find meanings of words in stories by image search.

## THE BLEEK AND LLOYD |XAM DICTIONARY

This digital publication is part of a Llarec project to digitise, research and publish the Bleek and Lloyd Archive. Llarec (the Lucy Lloyd Archive, Resource and Exhibition Centre) is a University of Cape Town research centre located at the Michaelis School of Fine Art.

The project has been made possible by funding provided by the Andrew W. Mellon Foundation and De Beers; and is the result of the cooperation of the three curating institutions: University of Cape Town, Iziko South African Museum and The National Library of South Africa.



THE BLEEK LLOYD  
|XAM DICTIONARY

[English-Ixam Dictionary](#)



THE BLEEK LLOYD  
|XAM DICTIONARY

[Ixam-English Dictionary](#)





# Bleek and Lloyd: Transcription

*Kyle Williams, MSc; Ngoni Munyaradzi, MSc*

- Using machine learning to transcribe |Xam.
- Training data manually generated.
- 45% accuracy at best.
  
- Crowdsourcing had 10% better performance.
  - Answer determined by agreement among 3 amateur transcribers.





# Bleek and Lloyd: Text Input

*Sunkanmi Olaleye, MSc*

- ❑ Inputting |Xam is non-trivial.
- ❑ Diacritics above, below and both; single and multiple characters.
- ❑ Custom Android keyboards for predictive and directed text entry in |Xam.





# Bantu Language IR

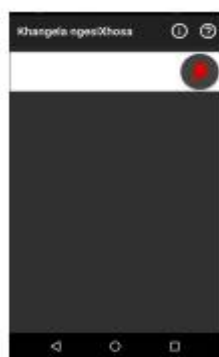
- Search engines in Bantu languages, especially South African languages (isiZulu, isiXhosa, etc.).
- Many core IR algorithms are unchanged but some language-specific algorithms needed:
  - Language identification
  - Text pre-processing and normalization
  - Ranking and reranking



# Bantu Language IR: Speech UI

*Morebodi Modise, MSc*

- Speech-driven mobile search interface in isiXhosa.
- Works well, but educated people want English!



(a) Query submission interface



(b) Detecting voice queries



(c) Detected voice query with list of voice results

Fig. 3: Mobile voice interface

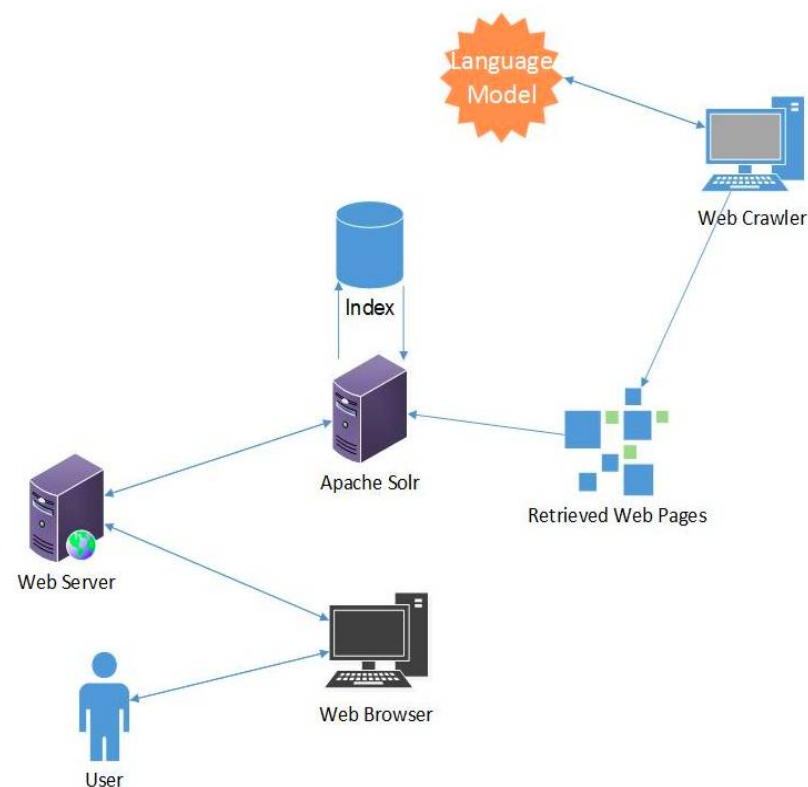


# Bantu Language IR: AfriWeb

*Nkosana Malumba, Katlego  
Moukangwe, BSc(Hons)*

- ❑ Zulu Search Engine.
- ❑ High accuracy in identifying isiZulu vs. English+Italian.
- ❑ Simple morphological parser outperformed simple stemmer in IR results.

System Overview





# Bantu Language IR: Transfer?

*Nyasha Katemauswa, U/G*

## ▣ Shona Search Engine.

- Can we adapt the isiZulu framework to get better results in chiShona?

*Michael Kyeyune, U/G*

## ▣ Xhosa Search Engine.

- Can we adapt the isiZulu framework to get better results in isiXhosa?





# Bantu Language IR: Learning

*Mantombi Manqele, U/G*

- ▣ SeSotho Search Engine.
  - Can we improve search results using word embeddings for query expansion?





# Bantu Language IR: Similar Language IR

*Catherine Chavula, PhD (current);  
Sinead Urisohn, Andre Lopes, BSc(Hons)*

- Exploit language similarity for those who can read multiple languages.
  - Reranking to emphasize language similarity in addition to relevance.
  - Universal language group text pre-processing, such as stemming.





# Bantu Language IR: kiSwahili

*Joseph Telemala, PhD (current)*

- How do we support Swahili speakers?
  - Professionals want English for work.
  - Everyone wants kiSwahili for play.
- Who you are and what you are doing dictates query/result expectations.





# Corpora

- Corpora for African Language IR are rare.
  - There are limited corpora for speech recognition, speech synthesis, MT, etc.
  
- Very few documents online.
- Wikipedia has <1000 (poor quality) pages in many Bantu languages!
- Lots of OOV, loan words, mixed texts, etc.





# Corpora: Language Detection

*Meluleki Dube, U/G*

- Can we successfully determine the language, from among a group of 9 related African languages, of a piece of text?
  - Web page?
  - Tweet?
- Trigram modelling and model alignment distance gives up to 92% accuracy.
  - Incorrect predictions scatter by language similarity.





# Corpora: Crowdsourcing

*Sean Packham, MSc*

- Parallel corpus in isiXhosa-English.
- Will people contribute if money paid is varied or there is no money but only gamification?
  - Payment is only criterion!







# Corpora: SALANG

*Andreas von Holy, Osher Shuman, Alon Bresler,  
Bsc(Hons)*

- Create a central portal for documents in any SA Bantu language, with gamification, multilingual search, etc.

The screenshot shows the BantuWeb website interface. At the top, there is a navigation bar with the site name 'BantuWeb', links for 'Browse' and 'Contribute', and user information for 'andreas' including 'Logout' and 'Language' options. The main content area is divided into several sections:

- Leaderboard:** A table showing user profiles with their usernames and points. The top user has 1110 points.
- Newest Users:** A table showing the most recent users with their usernames and the date they joined.
- Search and Gamification:** A central message states 'You need 265 points to reach level 9!' and 'Upload 5 or Review 18 documents and reach level 9!'. Below this is a search bar and a 'Search' button. A prompt encourages users to 'Start searching for content in your language now!'.
- Language Selection:** A section titled 'Select your primary language:' with radio buttons for isiNdebele, isiXhosa, isiZulu, Sesotho, Sepedi, Setswana, and eSwati. A secondary section asks users to 'Select the language(s) you would like to retrieve your results in (including your primary language):' with checkboxes for the same languages.
- Lastest Uploads:** A table listing recent document uploads with their titles and languages. Examples include 'Kubalulekile ...' (isiXhosa), 'Umbhizo wo...' (isiXhosa), 'Unyaka wokuga...' (isiXhosa), 'U vha murema' (Tshivenda), and 'uqhankqalazo ...' (isiXhosa).
- Some Statistics:** A table showing various site statistics and their counts: Uploads (178), Reviews (420), isiXhosa Resources (21260), Sepedi Resources (26725), and Xitsonga Resources (9841).





# Corpora: Long-term effects

*Jackson Moji, MSc (current)*

- Does gamification for corpus creation work in the long term?
  - Will people lose interest?
  - Will they continue to contribute?
  - How is intrinsic motivation affected by time?
  
- Extension of SALang project.





# IR for Development

*Gina Paihama, PhD (current)*

- ▣ How can we give users directed results to address unemployment?

*Selvas Mwanza, PhD (current)*

- ▣ Can we use Twitter data to evaluate developmental measures in society (e.g., level of free speech)?





# Low Cost Digital Library Systems

*Hussein Suleman (current)*

- How can we design digital archives for low maintenance and high probability of data preservation?
  - Offline archives
  - Offline and pre-indexing
  - No database/Web services
- Working with various heritage archives at UCT/Wits/etc. as test sites.





## Where we are: what we learnt

- Simple gains are possible with language knowledge.
  - NLP techniques are slowly evolving, but statistical techniques (detection, correction, etc.) work well.
- Core IR approaches mostly work.
  - With some African language-specific tools.
- People have complex multilingual profiles.
  - Does not match search engine assumptions.





# Where we are: the unsolved problems

- Text corpora
  - Crowdsourcing is not working.
  - Government intervention is very costly.
- 600 Bantu languages
  - We need generic solutions.
- Using search/data to drive development.
  - Harder than it seems ...





# Opportunities and AFIRM

- ❑ In January UCT hosted AFIRM: Africa Summer School on ML for Search and Data Mining  
<http://sigir.org/afirm2019>
- ❑ ACM SIGIR and SIGKDD want to develop research and researchers, and help to build networks.
- ❑ In the pipeline:
  - AFIRM 2020 at UCT; AFIRM 2021 in East Africa?
  - SIGIR 2019/2020 events
  - Formal competitive evaluation campaign to develop datasets and algorithms and support researchers
- ❑ *How can IBM Research collaborate?*



# questions, comments, ...



*<http://dl.cs.uct.ac.za/>*

*enkosi  
hamba kakuhle  
thank you and go well*