African Language Information Retrieval

Hussein Suleman

University of Cape Town
Department of Computer Science
Digital Libraries Laboratory @ Centre for ICT4D

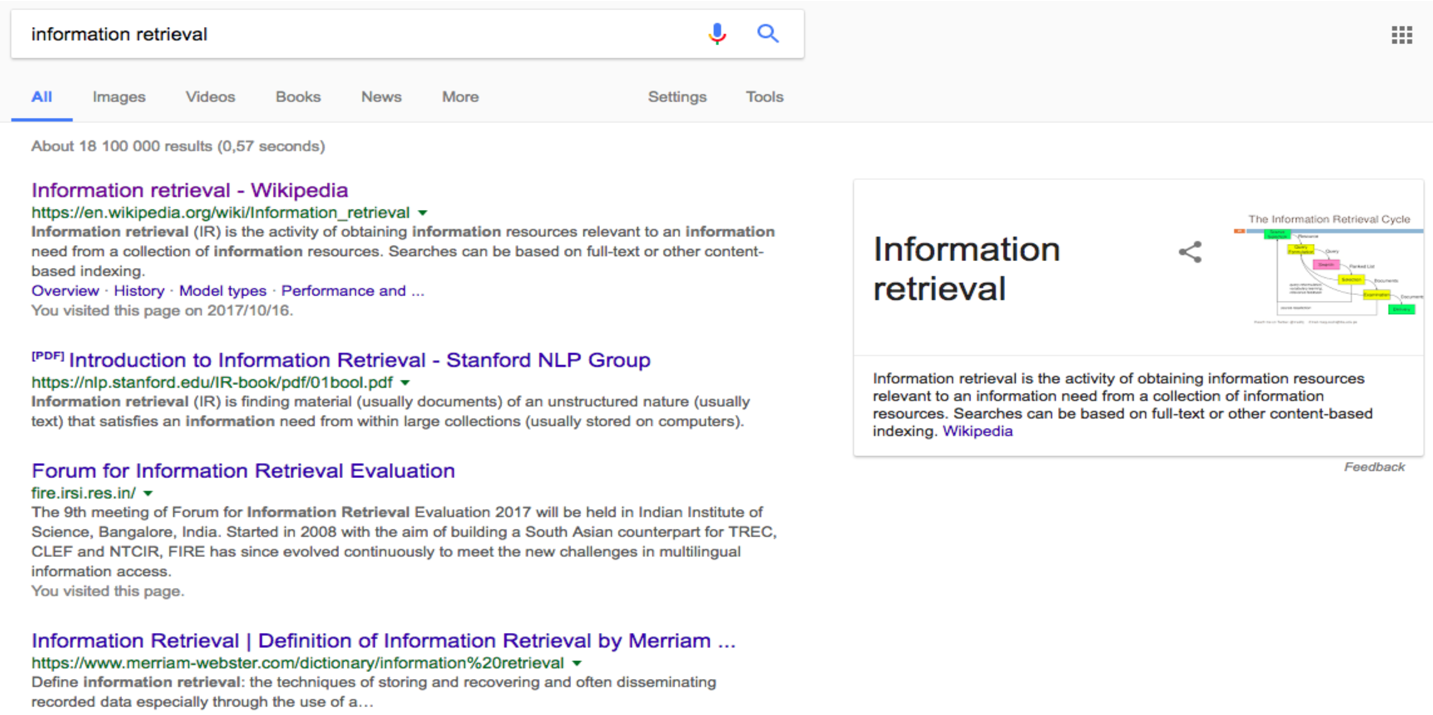
December 2017

Digital Libraries Lab

- Making information available to people
- Context-sensitive: low resource environment, different languages, different skills/culture, etc.
- 4 areas of interest:
  - Information Retrieval - HS
  - Educational Technology - HS
  - Knowledge Engineering - CMK
  - ICT4D Applications - various
- Some production systems / products

Information Retrieval (IR)

- Finding the most relevant information to satisfy an information need.



information retrieval

All Images Videos Books News More Settings Tools

About 18 100 000 results (0,57 seconds)

Information retrieval - Wikipedia
https://en.wikipedia.org/wiki/Information_retrieval ▾
Information retrieval (IR) is the activity of obtaining **information** resources relevant to an **information** need from a collection of **information** resources. Searches can be based on full-text or other content-based indexing.
Overview · History · Model types · Performance and ...
You visited this page on 2017/10/16.

[PDF] Introduction to Information Retrieval - Stanford NLP Group
https://nlp.stanford.edu/IR-book/pdf/01bool.pdf ▾
Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an **information** need from within large collections (usually stored on computers).

Forum for Information Retrieval Evaluation
fire.irsi.res.in/ ▾
The 9th meeting of Forum for **Information Retrieval** Evaluation 2017 will be held in Indian Institute of Science, Bangalore, India. Started in 2008 with the aim of building a South Asian counterpart for TREC, CLEF and NTCIR, FIRE has since evolved continuously to meet the new challenges in multilingual information access.
You visited this page.

Information Retrieval | Definition of Information Retrieval by Merriam ...
https://www.merriam-webster.com/dictionary/information%20retrieval ▾
Define **information retrieval**: the techniques of storing and recovering and often disseminating recorded data especially through the use of a…

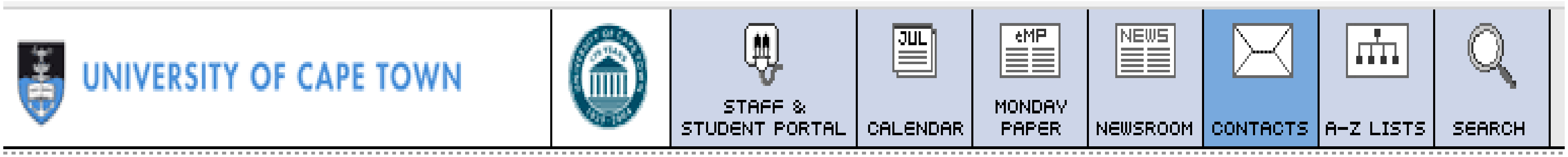
Information retrieval

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on full-text or other content-based indexing. Wikipedia

Feedback

IR in Africa

- Little understanding of local languages.
- Little relevant application.
- Poor general understanding of concept of search.
- 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.

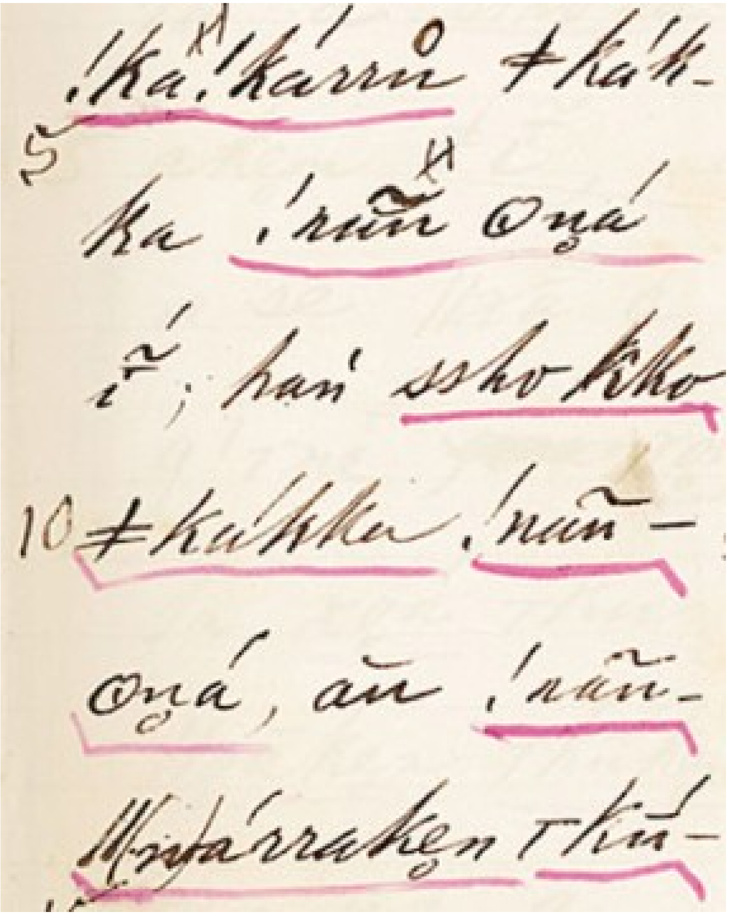
Mixed Language IR

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.

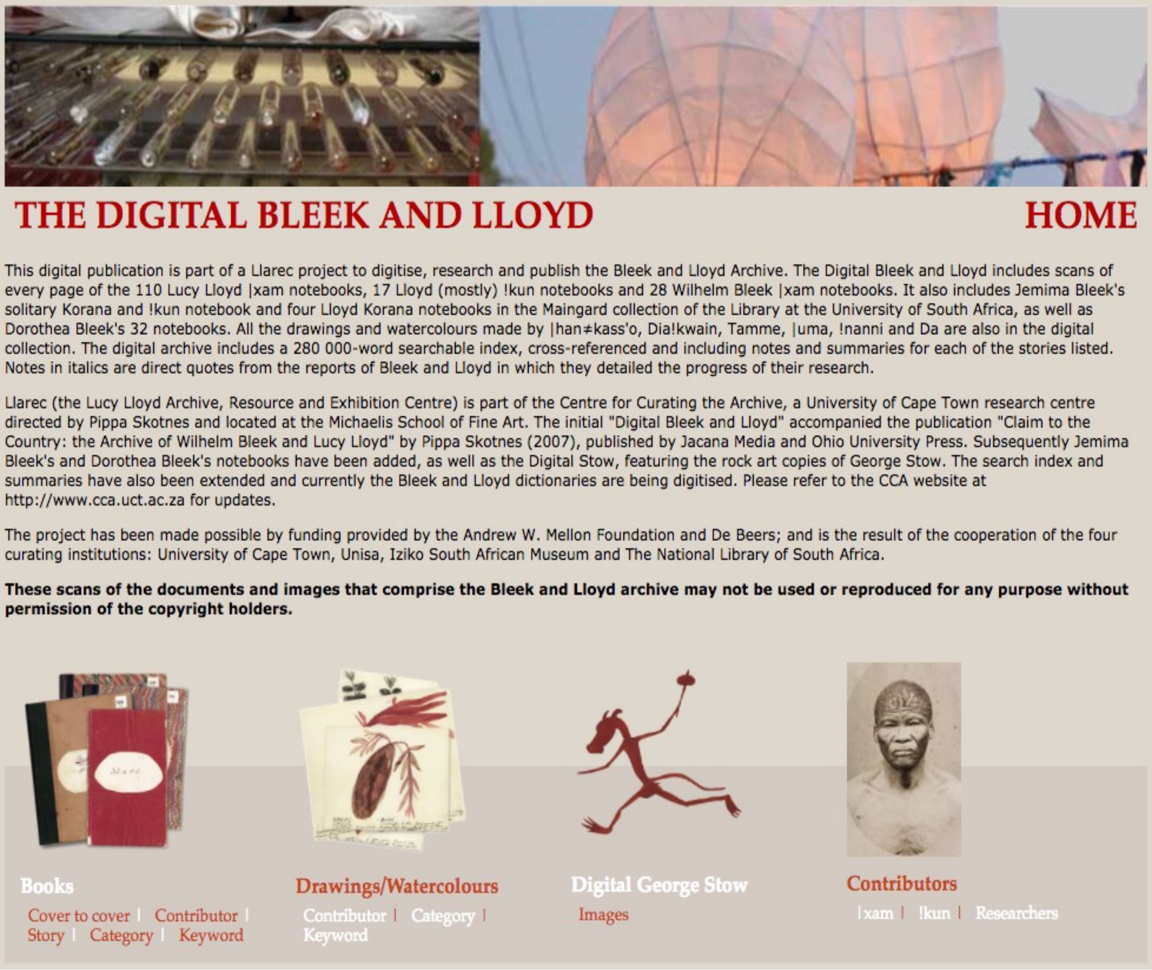
|Xam IR

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



!ka !karrū ǂkah-
5 ka !nūn ŏnā
ī; han' sshō kko
10 ǂkakka !nūn-
ŏnā, au !nūn-
||kyarraken ||ku-

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, |uma, !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.

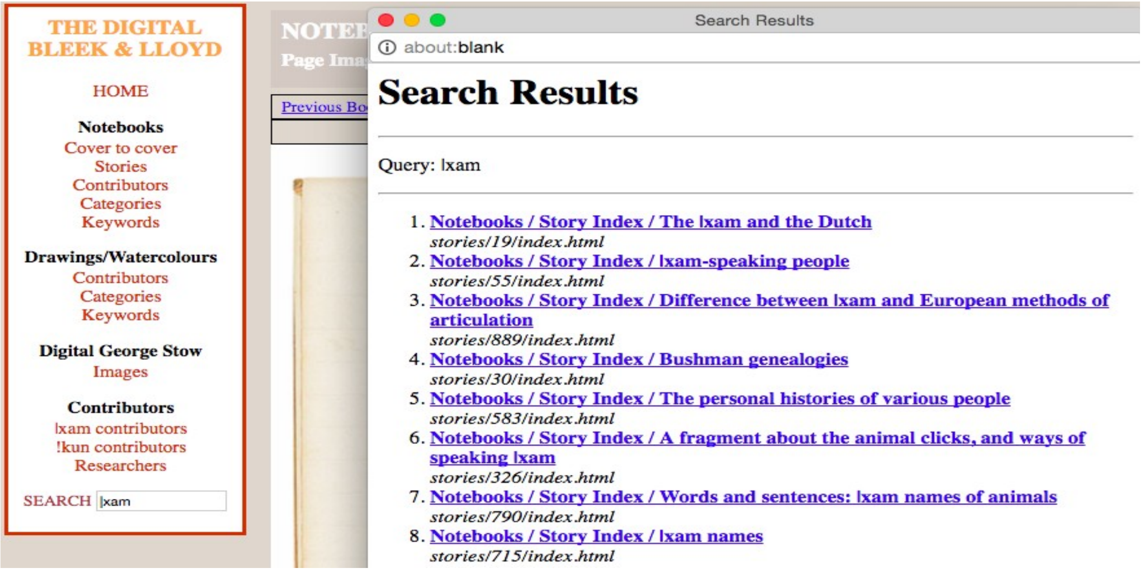
Llarec (the Lucy Lloyd Archive, Resource and Exhibition Centre) is part of the Centre for Curating the Archive, a University of Cape Town research centre directed by Pippa Skotnes and located at the Michaelis School of Fine Art. The initial "Digital Bleek and Lloyd" accompanied the publication "Claim 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.

Bleek and Lloyd: Low Resource IR

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



Search Results

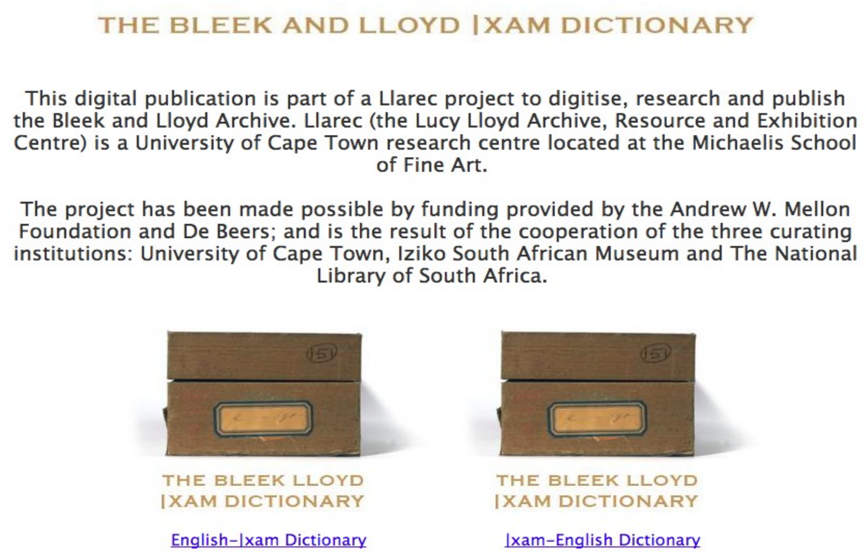
Query: |xam

1. Notebooks / Story Index / The |xam and the Dutch
   stories/19/index.html
2. Notebooks / Story Index / |xam-speaking people
   stories/55/index.html
3. Notebooks / Story Index / Difference between |xam 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 |xam
   stories/326/index.html
7. Notebooks / Story Index / Words and sentences: |xam names of animals
   stories/790/index.html
8. Notebooks / Story Index / |xam 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.

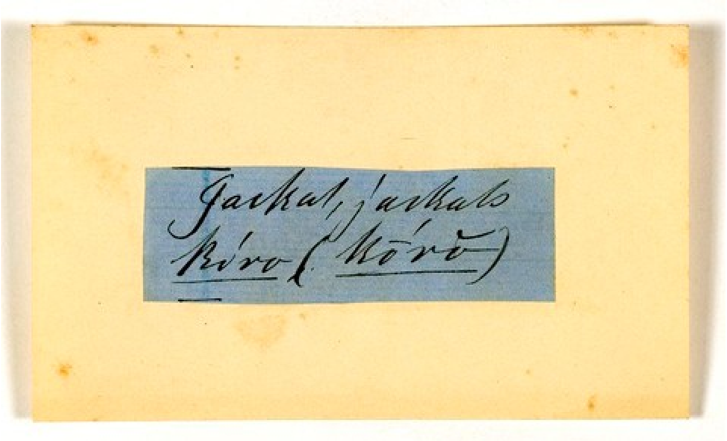
The project has been made possible by funding provided by the Andrew W. Mellon Foundation and De Beers; and is the result of the cooperation of the three curating institutions: University of Cape Town, Iziko South African Museum and The National Library of South Africa.

THE BLEEK LLOYD |XAM DICTIONARY

English–|xam Dictionary

THE BLEEK LLOYD |XAM DICTIONARY

|xam–English Dictionary



Jackal, jackals
kóro (kōrō)

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 (current)

- 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

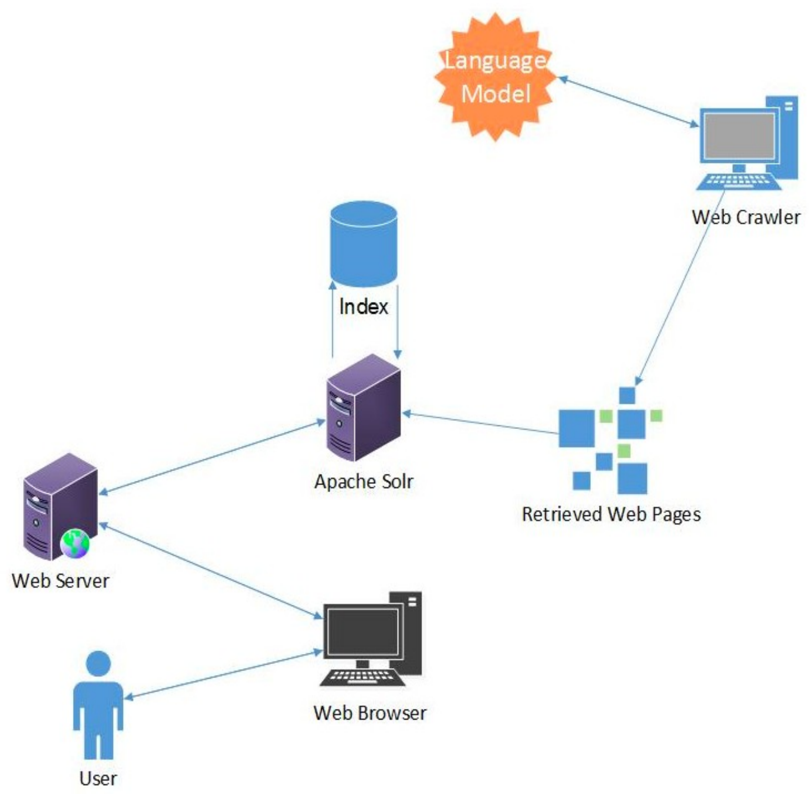
- Search engines in Bantu languages, especially South African languages (isiZulu, isiXhosa, etc.).
- Many core IR algorithms are unchanged but some language-specific algorithms needed:
  - Language identification
  - Text pre-processing and normalization
  - Ranking and reranking

Bantu Language IR: AfriWeb

Nkosana Malumba, Katlego Moukangwe, BSc(Hons)

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

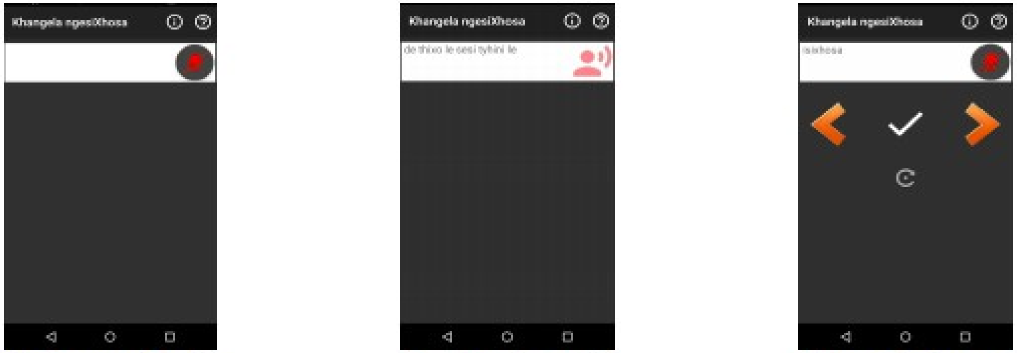
System Overview



Bantu Language IR: Speech UI

Michael Kyeyune, U/G; Morebodi Modise, MSc

- Speech-driven mobile search interface in isiXhosa.



(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: 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: Latest Projects

Nyasha Katemauswa, U/G

- ChiShona Search Engine.
  - Can we adapt the isiZulu framework to get better results in chiShona?
- KiSwahili IR.
  - Coming soon …

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 about 1000 (poor quality) pages in a large Bantu language collection!
- Lots of OOV, loan words, mixed texts, etc.

Corpora: Crowdsourcing

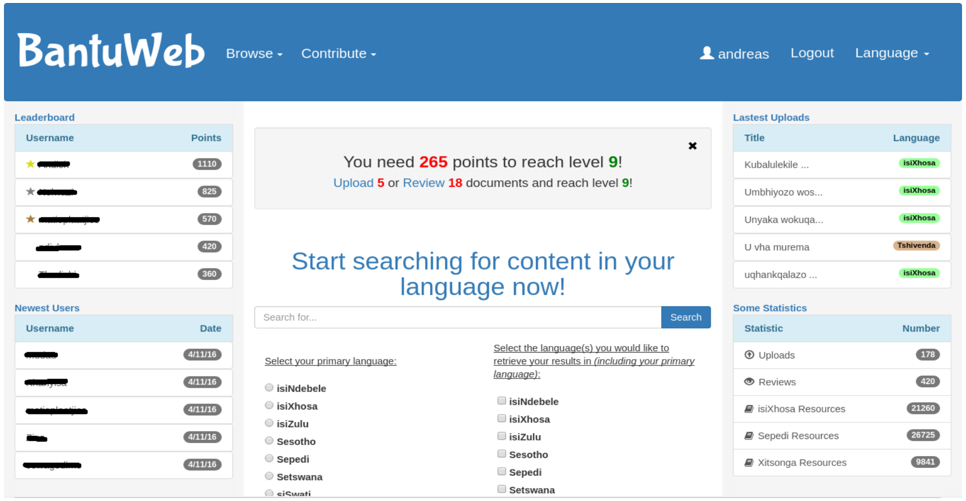
Sean Packham, MSc

- 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.



Leaderboard

Username	Points
★ [redacted]	1110
★ [redacted]	825
★ [redacted]	570
[redacted]	420
[redacted]	360

Newest Users

Username	Date
[redacted]	4/11/16
[redacted]	4/11/16
[redacted]	4/11/16
[redacted]	4/11/16
[redacted]	4/11/16

You need 265 points to reach level 9!
Upload 5 or Review 18 documents and reach level 9!

Start searching for content in your language now!

Search for... | Search

Select your primary language:

○ isiNdebele
○ isiXhosa
○ isiZulu
○ Sesotho
○ Sepedi
○ Setswana
○ siSwati

Select the language(s) you would like to retrieve your results in *(including your primary language):*

□ isiNdebele
□ isiXhosa
□ isiZulu
□ Sesotho
□ Sepedi
□ Setswana

Lastest Uploads

Title	Language
Kubalulekile ...	isiXhosa
Umbhiyozo wos...	isiXhosa
Unyaka wokuqa...	isiXhosa
U vha murema	Tshivenda
uqhankqalazo ...	isiXhosa

Some Statistics

Statistic	Number
Uploads	178
Reviews	420
isiXhosa Resources	21260
Sepedi Resources	26725
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)?

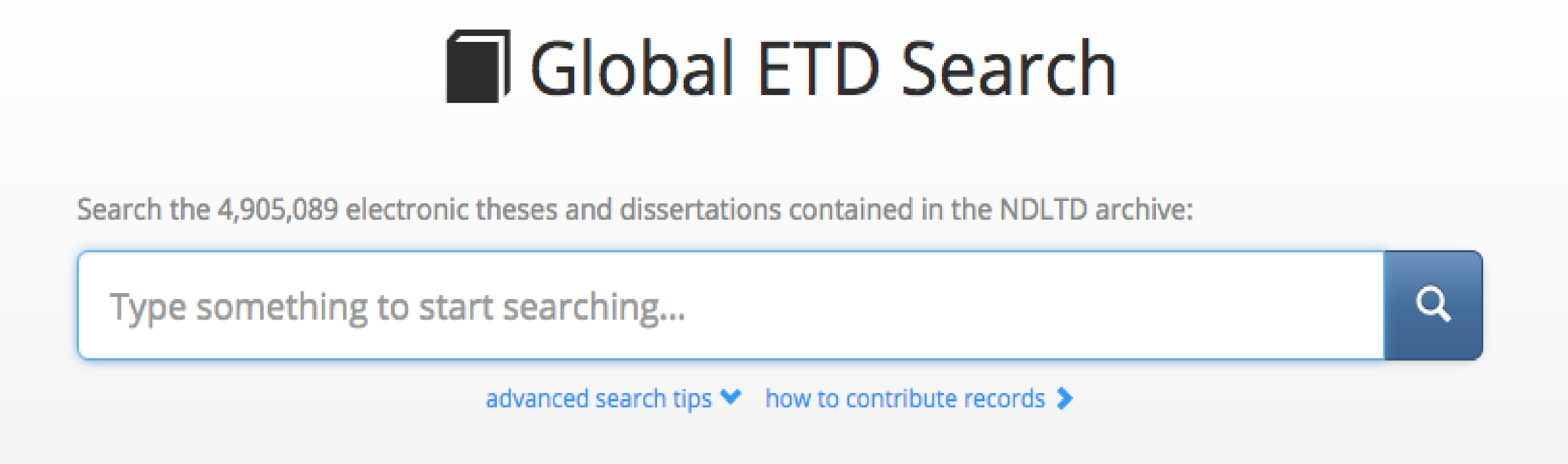
Other IR: Time-sensitive search

Jivashi Nagar, PhD (current)

- Can we exploit periodic patterns in search behaviour to provide contextually-relevant results?
  - Java in working hours leads to programming
  - Java over weekends leads to coffee

Other IR: ETD Metadata Search

NDLTD



Global ETD Search

Search the 4,905,089 electronic theses and dissertations contained in the NDLTD archive:

Type something to start searching...
advanced search tips
how to contribute records

questions, comments, ...

http://dl.cs.uct.ac.za/

enkosi
hamba kakuhle
thank you and go well