Information Retrieval in African Languages

Hussein Suleman

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

July 2018

Goal: IR for 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 State of Play in Africa

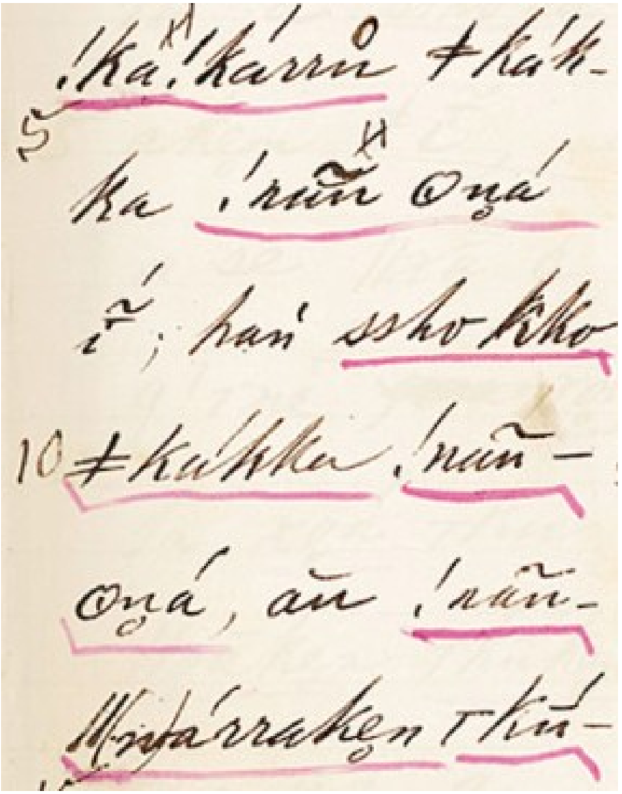
- Little algorithmic support in IR/NLP.
- Very little and noisy data.
  - <1000 Wikipedia documents
- Unclear language boundaries.
- Society not search savvy.
- Multilingualism.
- Mixed queries/documents.
- Resource limitations.

Phase One: Arabic

- isiZulu/isiXhosa understudied 10 years ago.
  - OTOH, Arabic spoken in lots of Africa, lots of interest and some documents.
- Most Africans mix languages so create mixed queries, and language-aware reranking provides better results.

Phase Two: |Xam

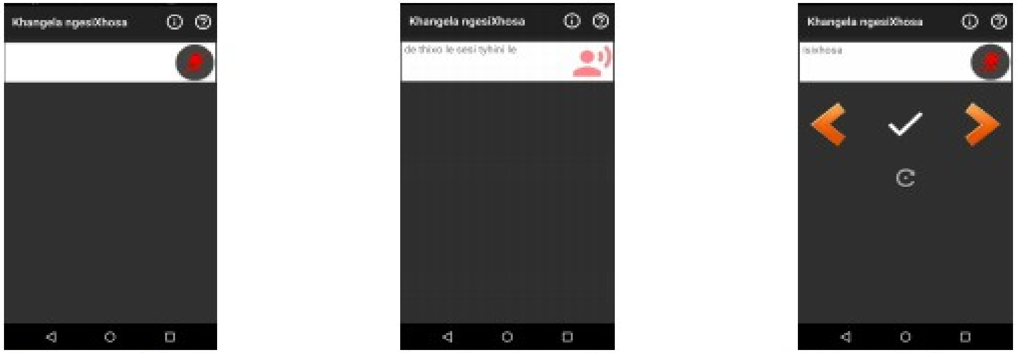
- Extinct Khoisan language.
- Language used in documenting early South African history/culture.
- No Unicode representation.
- Low-cost in-browser IR, visual search, machine-learnt/crowdsourced transcription.



!ka!karru ǂkakk-
5 ka ,nuñ ona
i ; han sshokkho
10 ǂkakka ,nuñ-
ona, au !nuñ-
||ŋiarraken rku-

Phase Three: Bantu Language IR

- Test systems in isiZulu, isiXhosa, and ciShona.
- Language identification (among 9 South African languages).
- Universal Bantu IR?
- Xhosa voice interface.



(a) Query submission interface

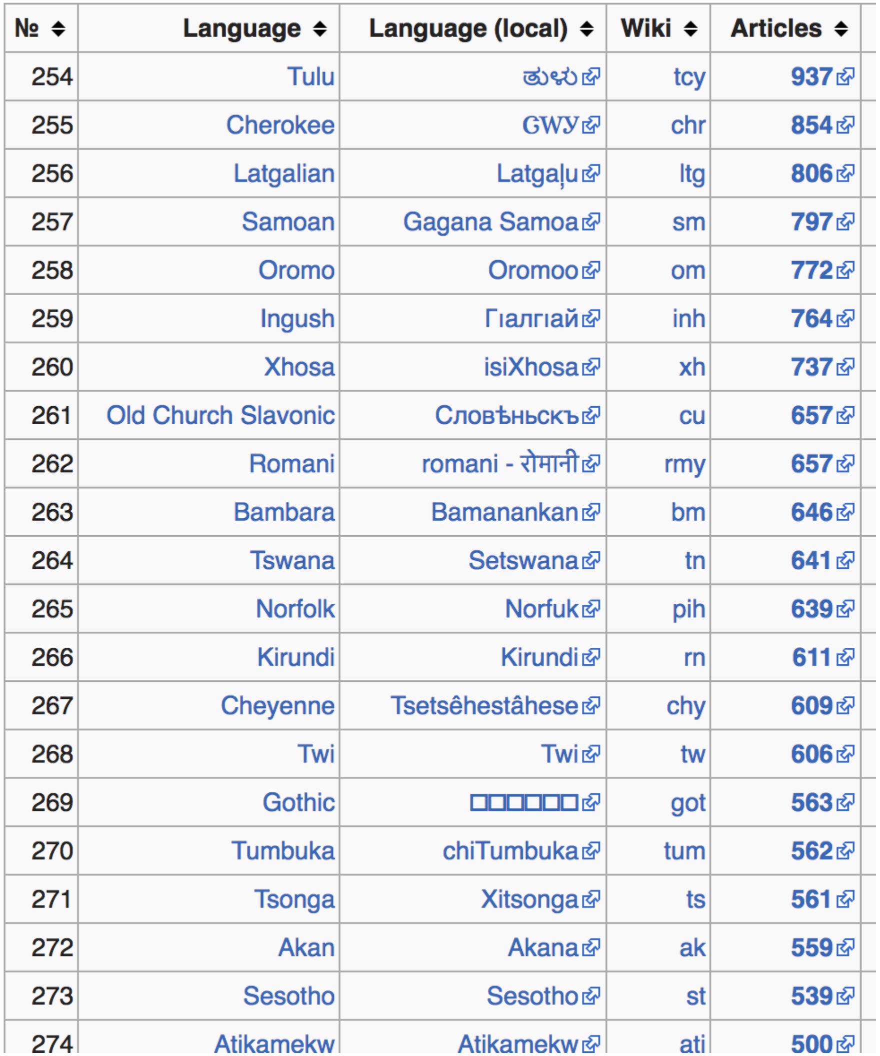
(b) Detecting voice queries

(c) Detected voice query with list of voice results

Fig. 3: Mobile voice interface

Corpora

- Crowdsourcing is not working.
  - Multi-language pooled approaches maybe?
- Intrinsic motivation?
  - Users want money!



№ ⬍	Language ⬍	Language (local) ⬍	Wiki ⬍	Articles ⬍
254	Tulu	ತುಳು	tcy	937
255	Cherokee	ᏣᎳᎩ	chr	854
256	Latgalian	Latgaļu	ltg	806
257	Samoan	Gagana Samoa	sm	797
258	Oromo	Oromoo	om	772
259	Ingush	Гӏалгӏай	inh	764
260	Xhosa	isiXhosa	xh	737
261	Old Church Slavonic	Словѣньскъ	cu	657
262	Romani	romani - रोमानी	rmy	657
263	Bambara	Bamanankan	bm	646
264	Tswana	Setswana	tn	641
265	Norfolk	Norfuk	pih	639
266	Kirundi	Kirundi	rn	611
267	Cheyenne	Tsetsêhestâhese	chy	609
268	Twi	Twi	tw	606
269	Gothic	□□□□□□	got	563
270	Tumbuka	chiTumbuka	tum	562
271	Tsonga	Xitsonga	ts	561
272	Akan	Akana	ak	559
273	Sesotho	Sesotho	st	539
274	Atikamekw	Atikamekw	ati	500

Where we are

- Some early successes but:
  - Too many languages, with
  - Too few documents,
  - Too few resources (money/users), and
  - Too much mixing of languages in queries and documents.

- Essentially, limited and noisy data ...

questions, comments, ...

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

enkosi
hamba kakuhle
thank you and go well