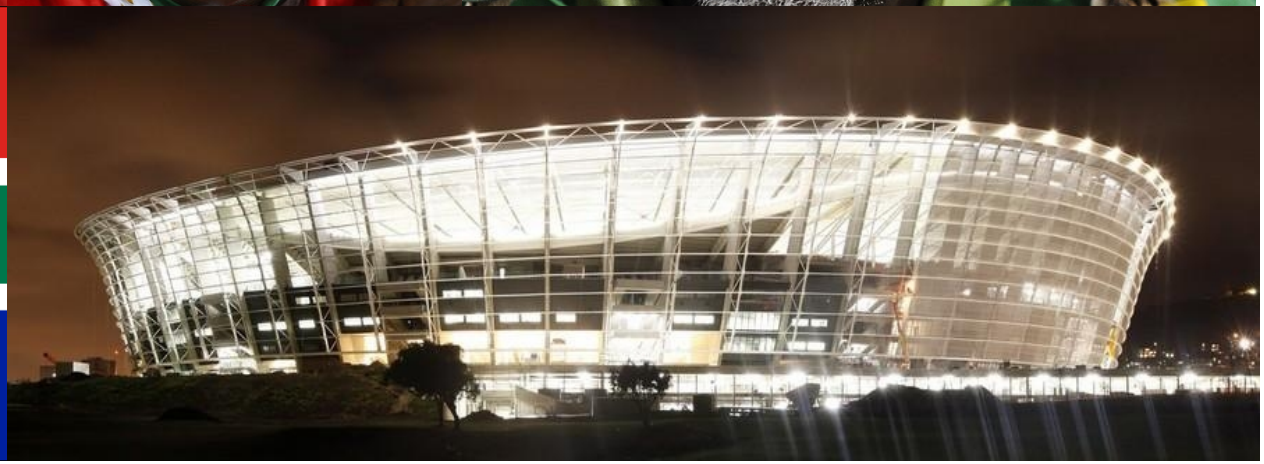
Cape Town - 2010 FIFA World Cup™ Host City

Ready to Welcome the World



ETD2011
SOUTH AFRICA

14th International Symposium on

Electronic Theses and Dissertations

September 2011 • Cape Town • South Africa

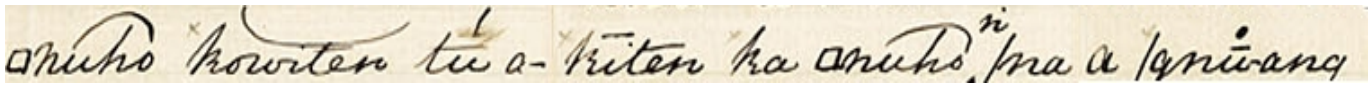
An international symposium and gathering of current and future researchers and practitioners in the area of Electronic Theses and Dissertations.

ETD 2011 will provide delegates with the unique opportunity to network, share experiences and discuss current good practices from around the world, enabling researchers and practitioners to chart a future course.

Translating Handwritten Bushman Texts

Kyle Williams and Hussein Suleman

Digital Libraries Laboratory
University of Cape Town



OUTLINE

- Bleek and Lloyd Collection
- Problem, motivation and solution
- Implementation
- Evaluation
- Conclusions

BLEEK AND LLOYD COLLECTION

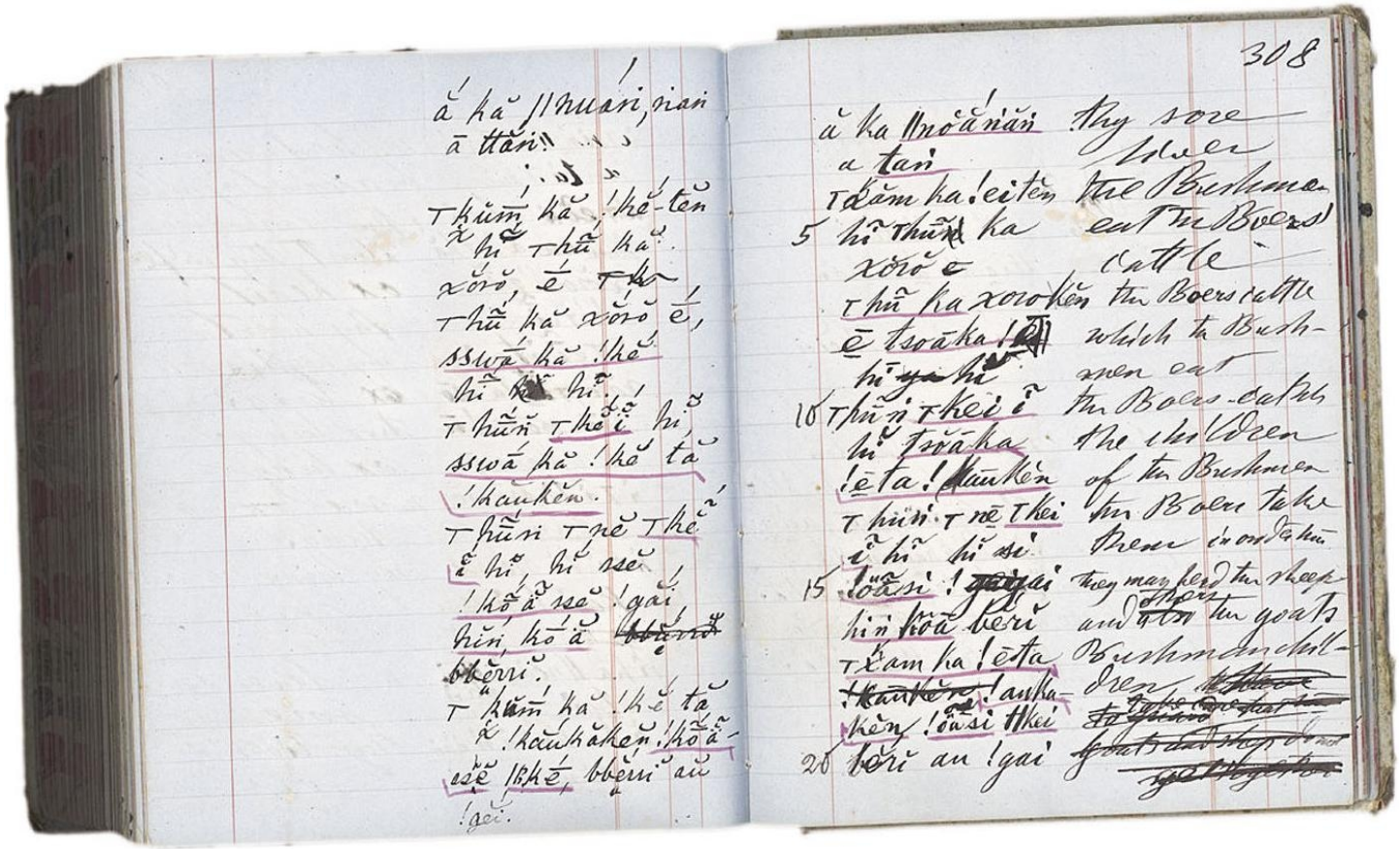
- Bushman people of Southern Africa
  - Earliest inhabitants of Earth
  - Unique view of the world
  - No living speakers of many Bushman languages

BLEEK AND LLOYD COLLECTION

- Collection contains notebooks, art and dictionaries
  - Bushman culture encoded in metaphorical stories
  - Preserving this collection → preserving Bushman culture

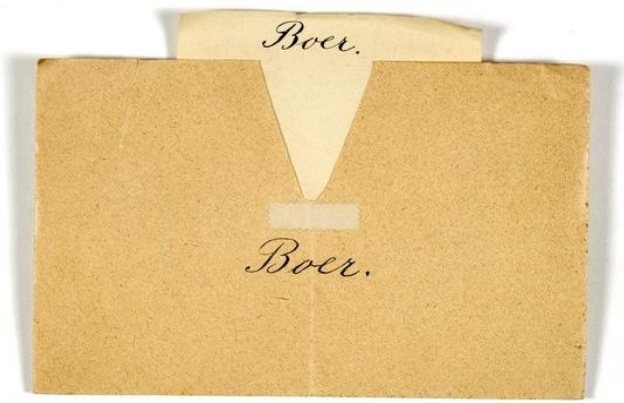
Digital Libraries Laboratory, University of Cape Town

BLEEK AND LLOYD COLLECTION

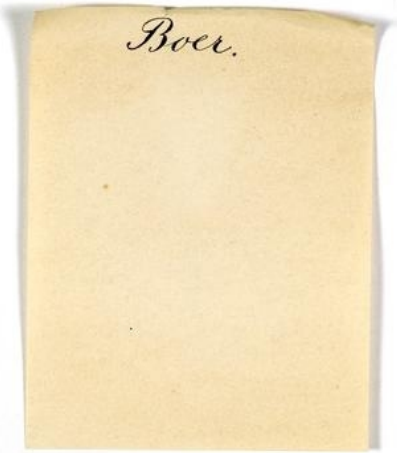


Digital Libraries Laboratory, University of Cape Town

BLEEK AND LLOYD COLLECTION



Envelope



Slip

Entry

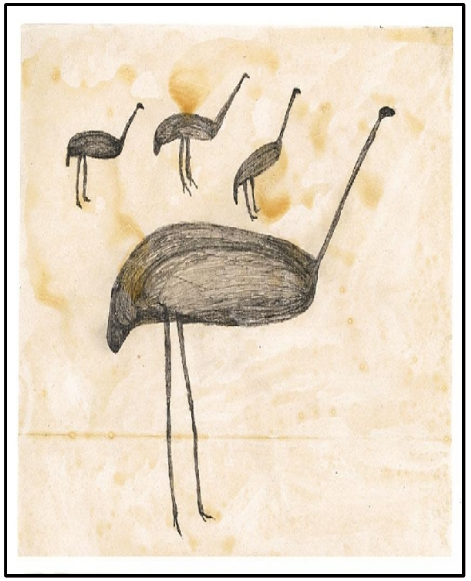
Digital Libraries Laboratory, University of Cape Town

MOTIVATION

- Collections have been digitised
- Systems have been built for preserving them
- Core services exist
- Next step involves digging into the text and build systems to assist with understanding

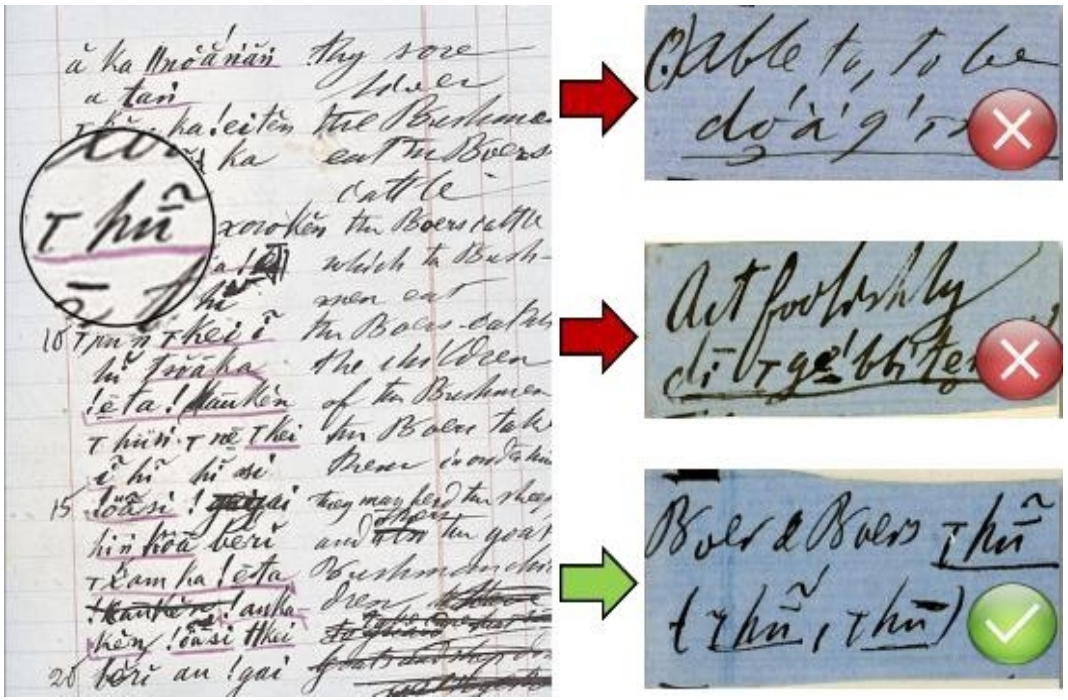
PROBLEM

- Notebooks contain information about Bushman language and culture
- Dictionary can be used by researchers to assist in understanding
- Manual translation impractical
  - Size of collection



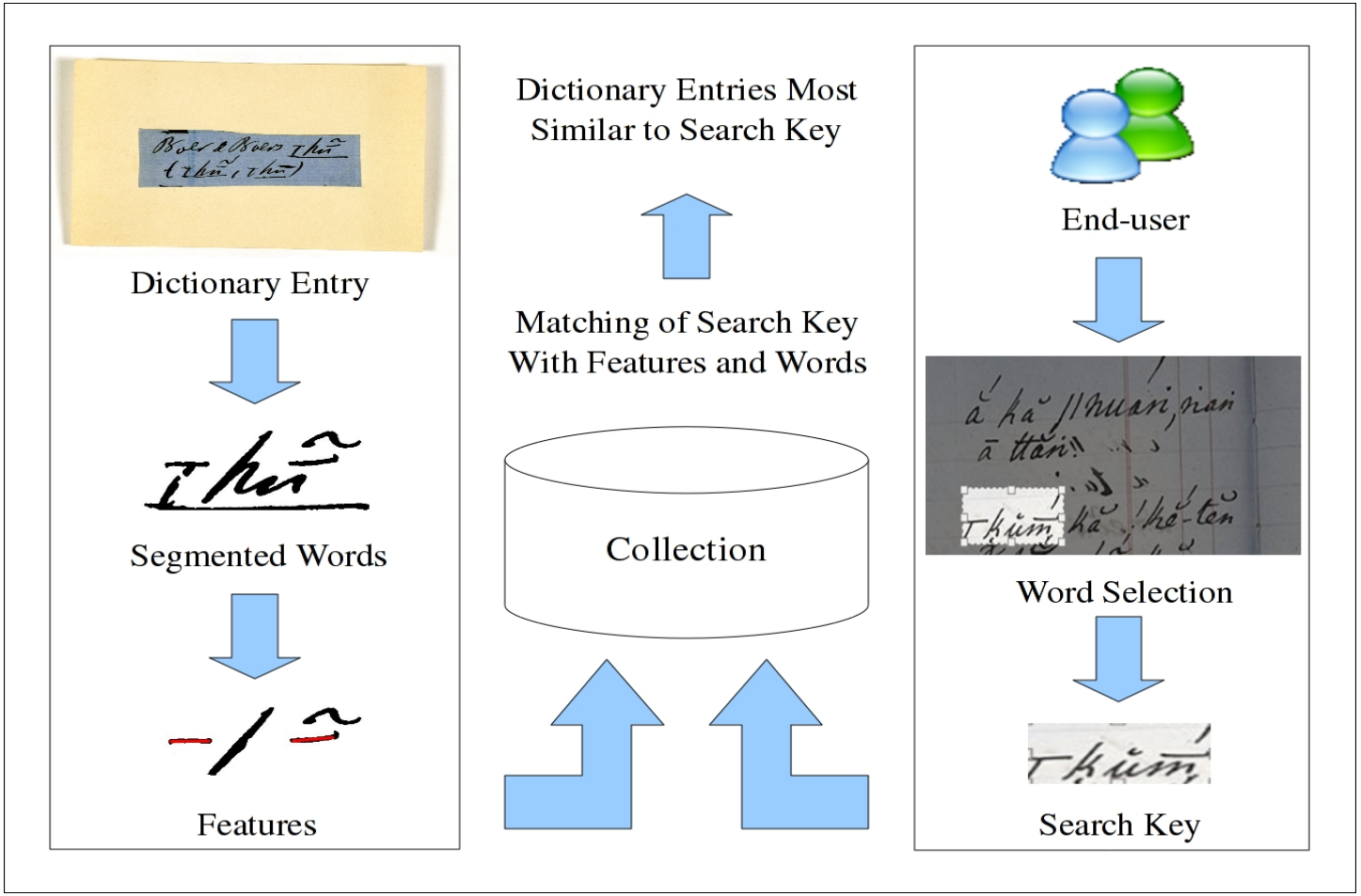
SOLUTION

- A system capable of returning a dictionary entry for a selected word in a notebook (CBIR)



Digital Libraries Laboratory, University of Cape Town

SYSTEM OVERVIEW



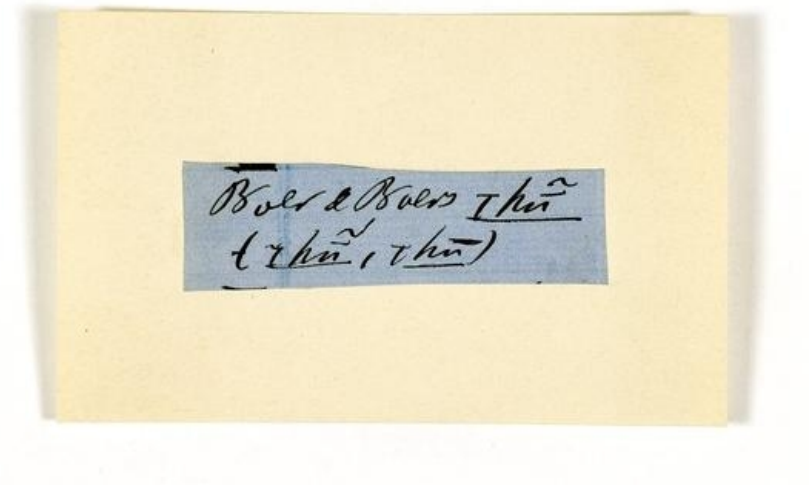
Digital Libraries Laboratory, University of Cape Town

IMPLEMENTATION

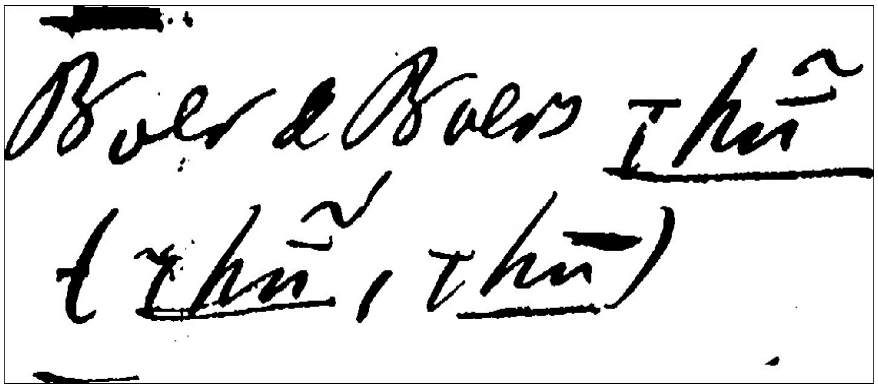
- Preprocessing
  - Image cleaning
  - Word segmentation
  - Feature extraction
- User input and matching
  - Key selection & setting variables
  - Feature matching → Accurate matching

PREPROCESSING

- Image Cleaning

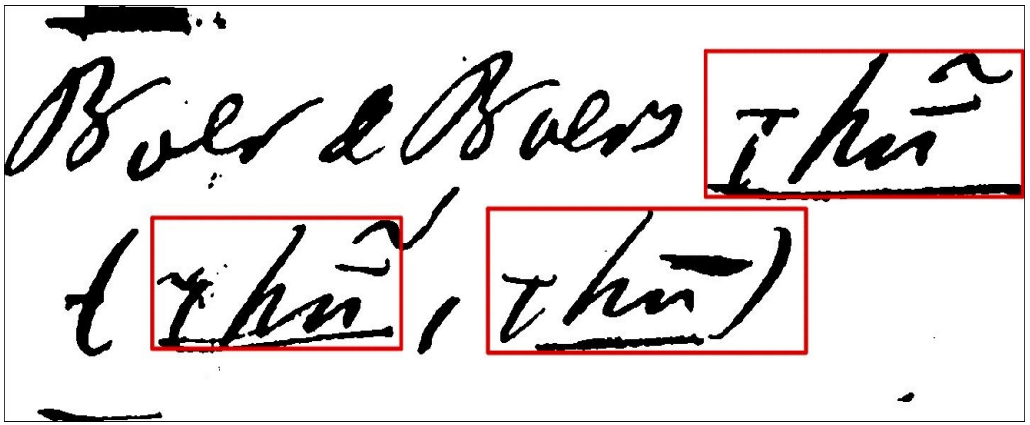


→



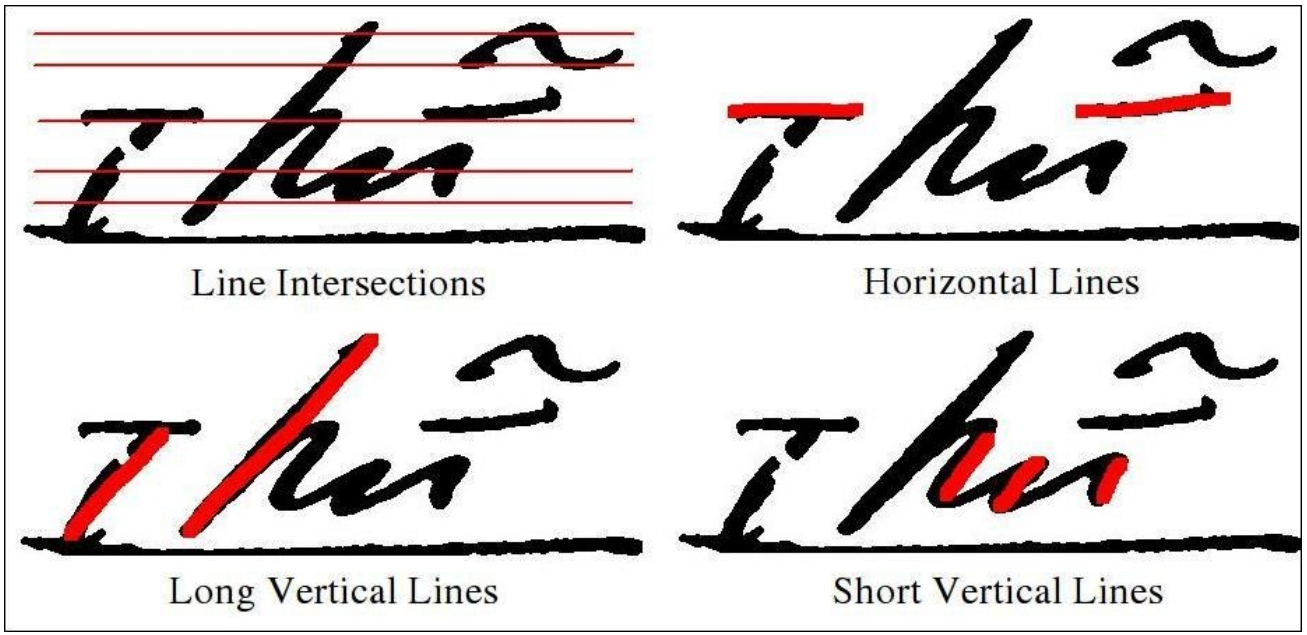
PREPROCESSING

- Word segmentation
  - Detect underlying lines (excludes English words)
  - Detect word boundaries



PREPROCESSING

- Feature extraction



FEATURE MATCHING

- Match words based on features
- Scores every word in collection based on feature similarity to search key
- Similar words will have a high feature score

FEATURE MATCHING

- Feature importance
  - Discriminatory power
- Variation
  - Allows for flexibility of matching features

- Return results above some threshold

ACCURATE MATCHING

- Three matching algorithms
  - DIF
  - XOR

Image 1 Image 2 XOR

  - Euclidean Distance Matching
- Return results above some threshold

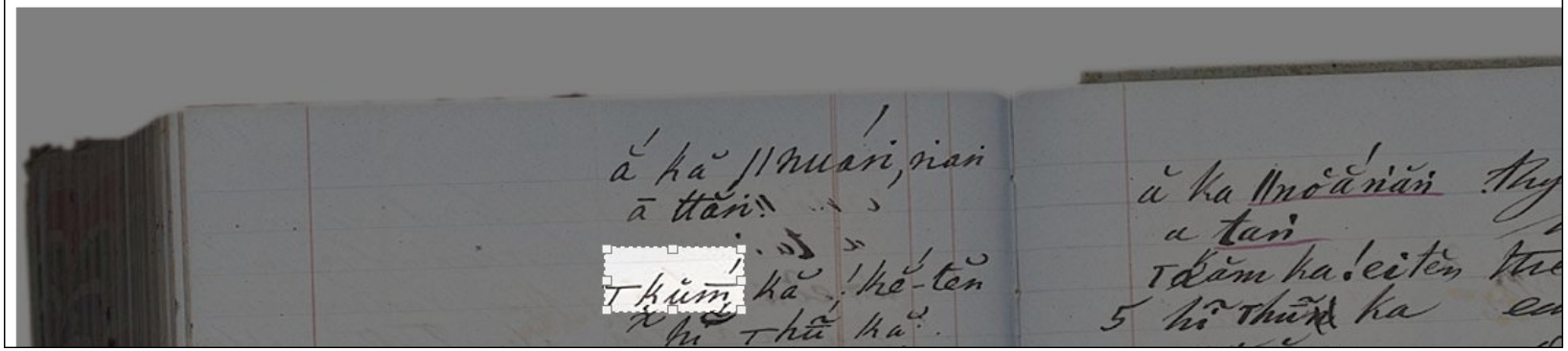
USER INPUT

BOLDProject :: Translator

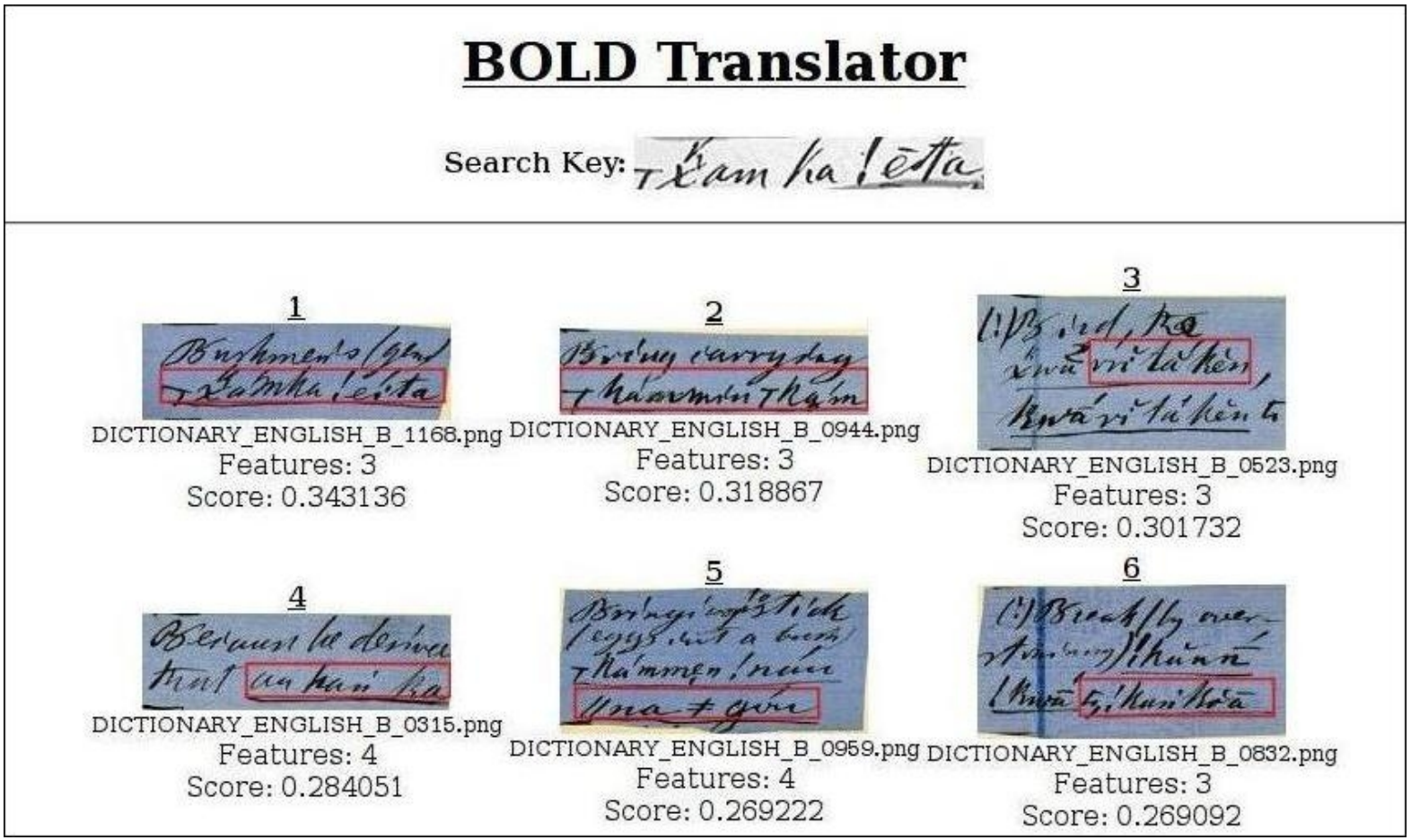
Highlight the word to translate and press the Translate! button

	Weight	Variation
Intersections:	25	0
Horizontal Lines:	25	1
Long Vertical Lines:	25	1
Short Vertical Lines:	25	1
Feature Threshold:		80 %
Matcher Threshold:		60 %
Matcher:	DIF	◉
	XOR	○
	EDM	○

Search!



RESULTS



Digital Libraries Laboratory, University of Cape Town

EVALUATION

- Each key selected 3 times

Key	Image	Size	Translation
Key 1		Small	Boer (farmer)
Key 2		Medium	Brother
Key 3		Large	Bushmen's gems

EVALUATION

- Segmentation was performed with 60% accuracy
- Feature Matching
  - Weights had little effect on results
  - Variation improved results
  - The best threshold was approximately 80%
  - Took 0.01 seconds for ~3000 images and 0.1 seconds for ~14000 image

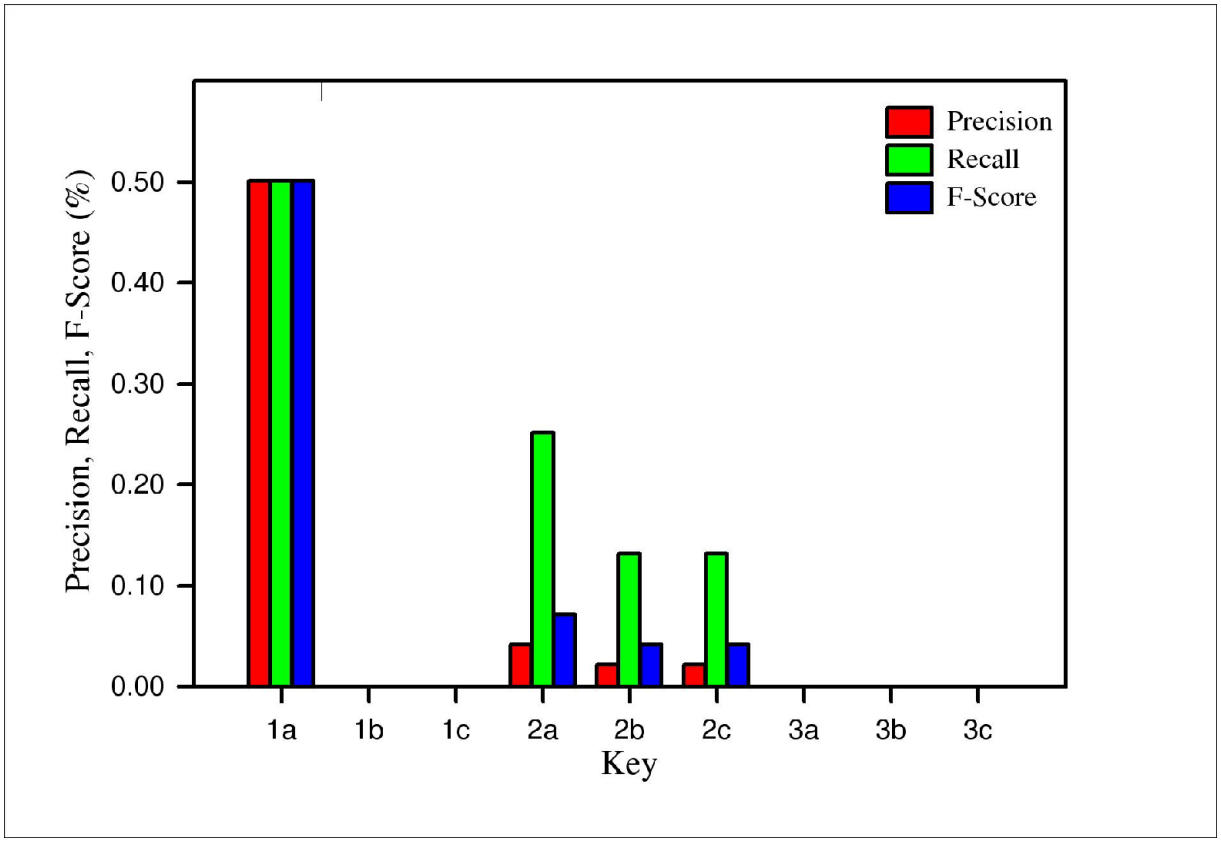
EVALUATION

- Accurate Matching
  - DIF algorithm was more accurate that XOR and EDM
  - DIF and XOR ran in approximately the same time while EDM was slow
  - Best threshold was approximately 60%

FULL SYSTEM EVALUATION

- 20% of collection ~3000 images
- Used optimal values obtained in previous experiments
  - Equal feature weights
  - Variation = 1
  - DIF Matching algorithm
  - 80% Feature threshold
  - 60% Matching threshold

FULL SYSTEM EVALUATION



Graph: Precision, Recall and F-score for end-to-end system

FULL SYSTEM EVALUATION

- Importance of well constrained key selection
- Recall remained mostly constant as scale increased while precision and F-score decreased
- System took ~1 second for 3000 images and ~16 seconds for 14000 images

CONCLUSIONS

- Built a system capable of matching words
- Returns positive results with good search keys
- Can be improved at all levels
- Could be applied to other collections
- Simple and efficient
- Can assist researchers in interpreting and understanding Bushman language and culture

THANK YOU

Questions?