ETD 2011
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
• 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

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

Digital Libraries Laboratory, University of Cape Town
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**
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

Dictionary Entry

Segmented Words

Features

Dictionary Entries Most Similar to Search Key

Matching of Search Key With Features and Words

Collection

End-user

Word Selection

Search Key

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

Digital Libraries Laboratory, University of Cape Town
PREPROCESSING

- Image Cleaning
PREPROCESSING

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

Digital Libraries Laboratory, University of Cape Town
PREPROCESSING

- Feature extraction

Line Intersections

Horizontal Lines

Long Vertical Lines

Short Vertical Lines

Digital Libraries Laboratory, University of Cape Town
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

Digital Libraries Laboratory, University of Cape Town
ACCURATE MATCHING

- Three matching algorithms
  - DIF
  - XOR

- Euclidean Distance Matching
- Return results above some threshold

Digital Libraries Laboratory, University of Cape Town
# BOLDProject :: Translator

Highlight the word to translate and press the Translate! button

<table>
<thead>
<tr>
<th></th>
<th>Weight</th>
<th>Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersections:</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>Horizontal Lines:</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Long Vertical Lines:</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td>Short Vertical Lines:</td>
<td>25</td>
<td></td>
</tr>
<tr>
<td>Feature Threshold:</td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td>Matcher Threshold:</td>
<td></td>
<td>60%</td>
</tr>
<tr>
<td>Matcher:</td>
<td>DIF</td>
<td></td>
</tr>
<tr>
<td></td>
<td>XOR</td>
<td></td>
</tr>
<tr>
<td></td>
<td>EDM</td>
<td></td>
</tr>
</tbody>
</table>

[Search!]

---

Digital Libraries Laboratory, University of Cape Town
RESULTS

BOLD Translator

Search Key: [Image of handwritten text]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="DICTIONARY_ENGLISH_B_1168.png" alt="Image" /></td>
<td><img src="DICTIONARY_ENGLISH_B_0944.png" alt="Image" /></td>
<td><img src="DICTIONARY_ENGLISH_B_0623.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>Features: 3</td>
<td>Features: 3</td>
<td>Features: 3</td>
</tr>
<tr>
<td></td>
<td>Score: 0.343136</td>
<td>Score: 0.318867</td>
<td>Score: 0.301732</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><img src="DICTIONARY_ENGLISH_B_0315.png" alt="Image" /></td>
<td><img src="DICTIONARY_ENGLISH_B_0959.png" alt="Image" /></td>
<td><img src="DICTIONARY_ENGLISH_B_0832.png" alt="Image" /></td>
</tr>
<tr>
<td></td>
<td>Features: 4</td>
<td>Features: 4</td>
<td>Features: 3</td>
</tr>
<tr>
<td></td>
<td>Score: 0.284051</td>
<td>Score: 0.269222</td>
<td>Score: 0.269092</td>
</tr>
</tbody>
</table>

Digital Libraries Laboratory, University of Cape Town
## EVALUATION

- Each key selected 3 times

<table>
<thead>
<tr>
<th>Key</th>
<th>Image</th>
<th>Size</th>
<th>Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key 1</td>
<td><img src="image1.png" alt="Image" /></td>
<td>Small</td>
<td>Boer (farmer)</td>
</tr>
<tr>
<td>Key 2</td>
<td><img src="image2.png" alt="Image" /></td>
<td>Medium</td>
<td>Brother</td>
</tr>
<tr>
<td>Key 3</td>
<td><img src="image3.png" alt="Image" /></td>
<td>Large</td>
<td>Bushmen’s gems</td>
</tr>
</tbody>
</table>

Digital Libraries Laboratory, University of Cape Town
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 images
EVALUATION

• Accurate Matching
  • DIF algorithm was more accurate than 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

Digital Libraries Laboratory, University of Cape Town
Graph: Precision, Recall and F-score for end-to-end system

Digital Libraries Laboratory, University of Cape Town
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

Digital Libraries Laboratory, University of Cape Town
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?