Using a Hidden Markov Model to Transcribe Handwritten Bushman Texts

Kyle Williams and Hussein Suleman
Department of Computer Science, University of Cape Town

Introduction

The automatic transcription of the Bushman texts in the Bleek and Lloyd Collection [3] is difficult due to the complex diacritics that appear in the text. In a small subset of the collection, 336 character/diacritic combinations were found. This poster describes an investigation into automatic recognition of Bushman text lines using a common Hidden Markov Model-based approach. Text lines were automatically segmented and supervised machine learning took place. The system was then evaluated by conducting a number of experiments that treated the diacritics in different ways.

Encoding and Text Capture

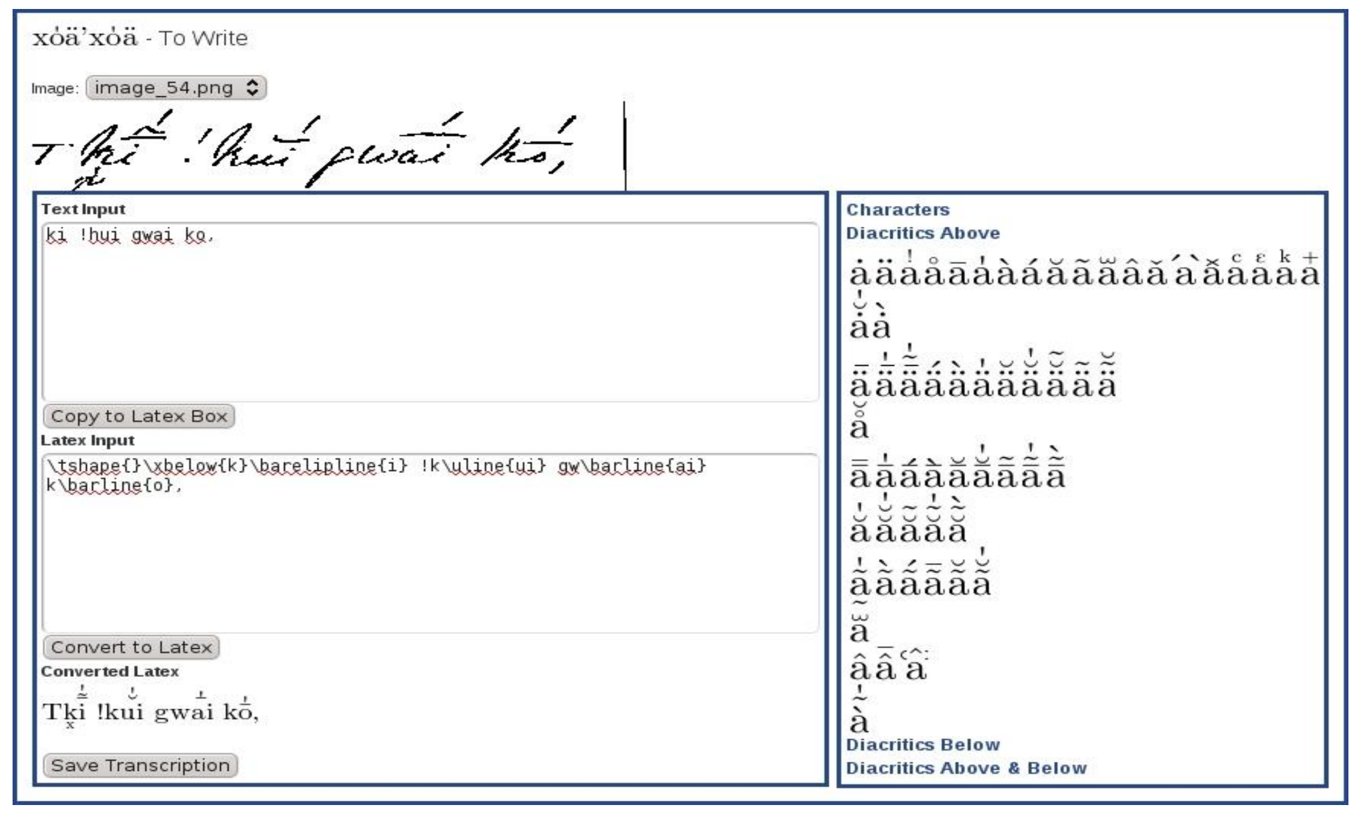

xȍä’xȍä - To Write

Image: image_54.png

Text Input

ki !hui gwai ko,

Latex Input

```
\tshape{}\xbelow{k}\bareliplne{i} !k\uline{ui} gw\barline{ai}
k\barline{o},
```

Converted Latex

Tki !kui gwai kō,

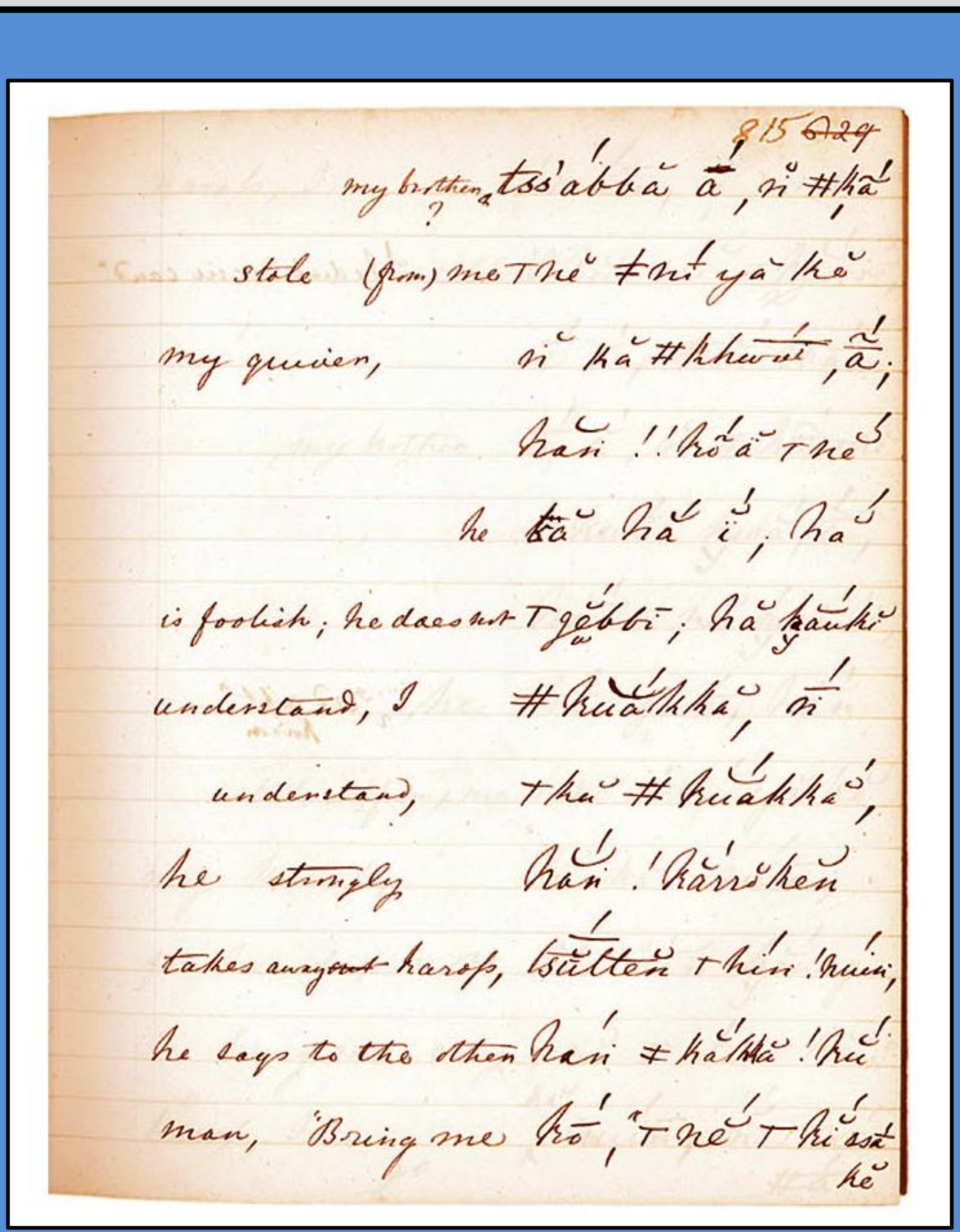
The Bushman text was represented using LaTeX and the TIPA package since it cannot be represented in non-custom Unicode. An AJAX tool was built to allow for the capturing of Bushman text by typing the base characters and adding diacritics.

References

[1] U. Marti and H. Bunke. On the influence of vocabulary size and language models in unconstrained handwritten text recognition. In ICDAR '01: Proceedings of the Sixth International Conference on Document Analysis and Recognition, page 260. 2001.

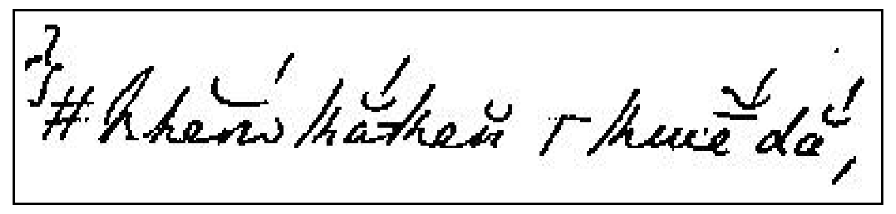
[2] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. Computers, IEEE Transactions on, C-23(1):90 -93, 1974.

[3] Hussein Suleman. Digital libraries without databases: The Bleek and Lloyd collection. In Research and Advanced Technology for Digital Libraries, pages 392-403. 2007



Preprocessing and Recognition

Bushman text was manually separated from English text and then thresholded using an adaptive thresholding [4] approach and text lines were automatically segmented using the horizontal projection profile [1].



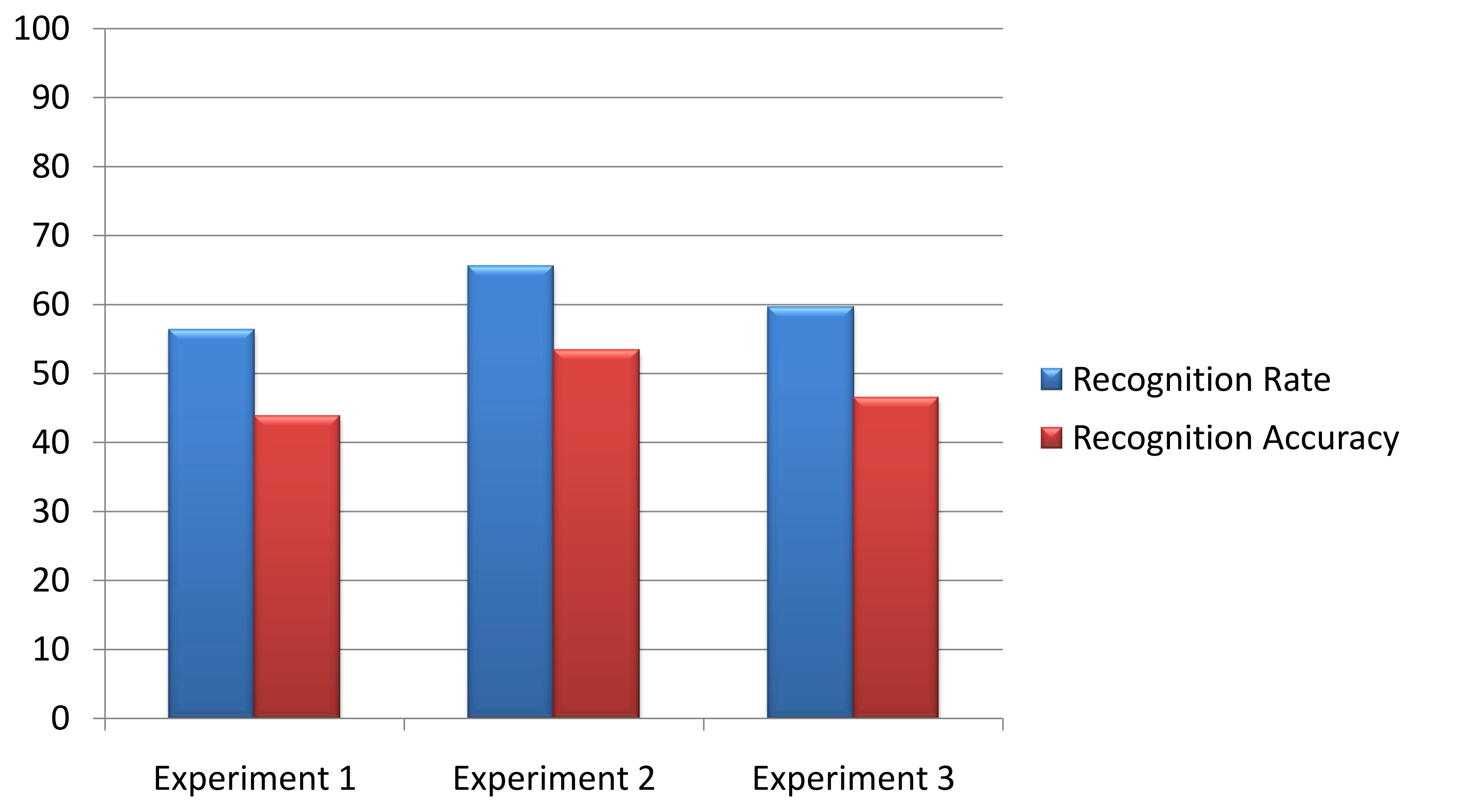
A HMM with 12 emitting states was built. Features were extracted using a sliding window approach and 9 geometrical features [1] as well as the coefficients from the Discrete Cosine Transform [2] were used. No statistical language model exists for the Bushman languages and therefore statistical information was not integrated into the recognition system.

Evaluation and Results

Evaluation was performed using 10-fold cross validation for 698 text-lines.

Recognition Rate = $\frac{N-S-D}{N}$ Recognition Accuracy = $\frac{N-S-D-I}{N}$

Experiment 1: Trained each character/diacritic combination and recognised them.
Experiment 2: Trained each character/diacritic combination and recognised ignoring diacritics. 10% of base characters correct but not diacritics.
Experiment 3: Trained and recognised characters ignoring diacritics. Indicates that multilayer classification may not be viable



Conclusions

Text lines were recognised using a HMM and, for the majority of cases, there are fewer than 3 training samples for each character/diacritic combination. An improvement could be achieved when diacritics were ignored, which is not viable for the Bushman languages. The findings suggest that the other techniques for recognition of Bushman languages need to be investigated.

This research is supported by the University of Cape Town, the Telkom/NSN/ Telesciences/THRIP Center of Excellence and the National Research Foundation.