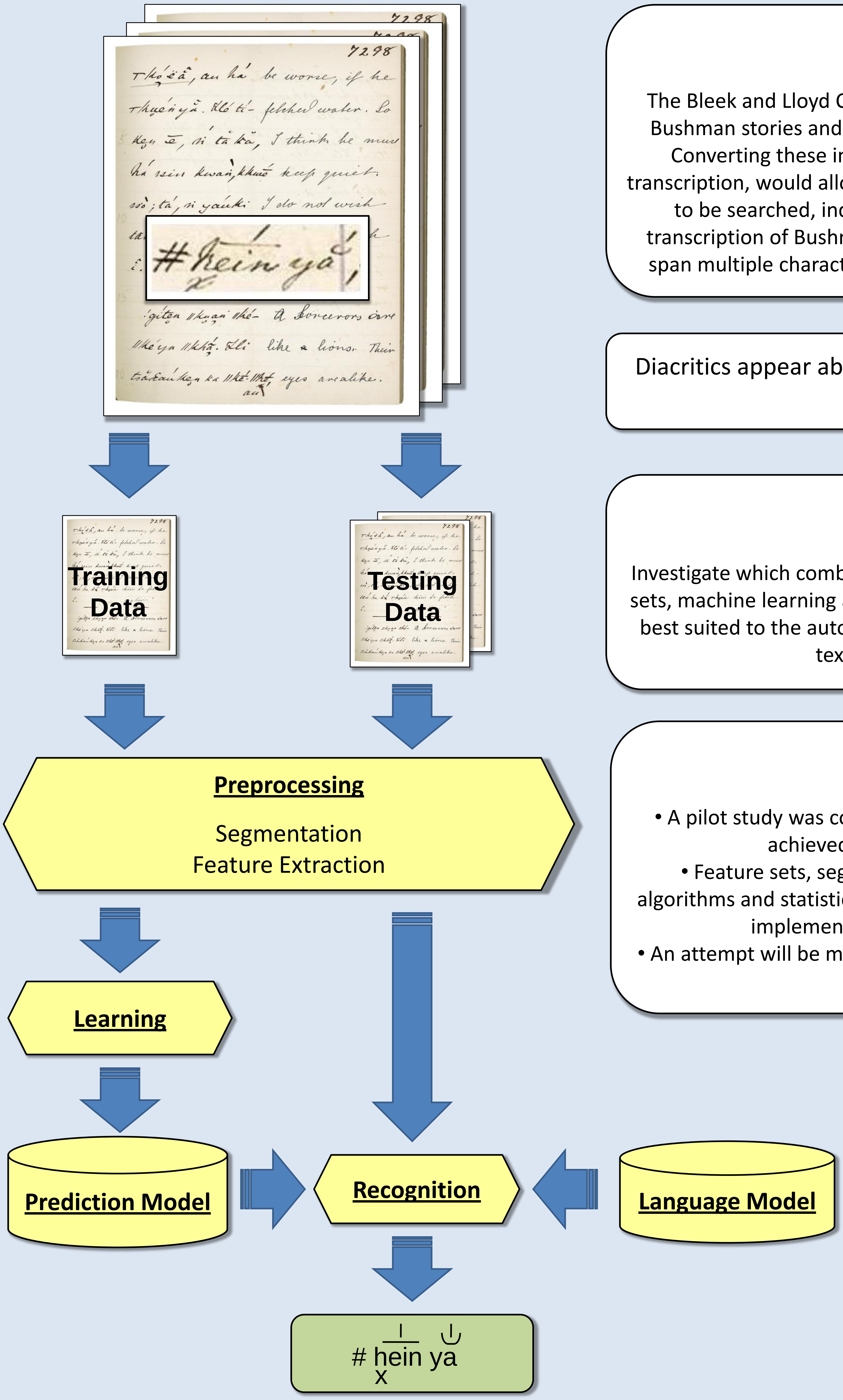
Learning to Read Bushman

Kyle Williams and Hussein Suleman



Training Data

Testing Data

Preprocessing

Segmentation
Feature Extraction

Learning

Prediction Model

Recognition

Language Model

# hein ya
x

Introduction

The Bleek and Lloyd Collection contains notebooks that document Bushman stories and culture and are available as digital scans [1]. Converting these images to text, through a process known as transcription, would allow for the text which appears in the notebooks to be searched, indexed and compared. However, automatic transcription of Bushman texts is difficult due to the diacritics that span multiple characters and appear above and below characters.

Diacritics appear above and below characters and also span multiple characters.

Objectives

Investigate which combinations of segmentation approaches, feature sets, machine learning algorithms and statistical language models are best suited to the automatic transcription of handwritten Bushman texts with complex diacritics.

Methodology

- A pilot study was conducted for neatly rewritten characters and achieved 80% transcription accuracy [2].
- Feature sets, segmentation approaches, machine learning algorithms and statistical language models from the literature will be implemented and evaluated experimentally.
- An attempt will be made to explain results within the framework of information theory.

Conclusions

This research will provide insight into different techniques for the automatic transcription of handwritten Bushman texts. The findings will also be applicable to other languages, especially those with complex diacritics.

References

[1] H. Suleman. Digital libraries without databases: The bleek and lloyd collection. In 11th European Conference on Research and Advanced Technology for Digital Libraries. March 2007.

[2] K. Williams. Feasibility of automatic transcription of neatly rewritten bushman texts. Technical report, Department of Computer Science, University of Cape Town, 2010. Technical Report CS10-06-00.

Acknowledgements

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

Further Information

Email:
kwilliams@cs.uct.ac.za
Web:
http://people.cs.uct.ac.za/~kwilliams
The Digital Bleek and Lloyd Collection:
http://lloydbleekcollection.cs.uct.ac.za