Learning to Read Bushman
MSc Research Proposal

Kyle Williams
Supervisor: A/Prof. Hussein Suleman
OUTLINE

• Introduction
• Project Description
• Pilot Study
• Research Questions & Evaluation
• Work Detail
• Anticipated Outcome
• Questions
INTRODUCTION

- Bleek and Lloyd Collection
- Cannot index, search or compare texts
- Need to transcribe text
  - Tedious and time consuming
- Need to automatically transcribe text
PROBLEM

- Diacritics & segmentation
- Diacritics and recognition
- English & Bushman text
- Lack of language experts
- No Bushman dictionary
- No language model
- Cannot be represented in Unicode
PROJECT DESCRIPTION

• Build a system capable of automatically transcribing handwritten Bushman texts
• Segmentation
• Feature extraction
• Machine Learning
  • HMMs, NNs and SVMs
PILOT STUDY

- Neatly handwritten text
- Limited character set
- Two authors
- 80% transcription accuracy using SVM
- Simplified problem but suggested feasibility
How accurately can the automatic transcription of handwritten Bushman texts be performed?

- Will be answered by addressing 3 sub-questions
Which of a selection of Hidden Markov Models, Neural Networks and Support Vector Machines, when used in conjunction with various feature sets, performs best when automatically transcribing handwritten Bushman texts?

- Use HMMs, NNs and SVMs with various feature sets
- Compare accuracy, word error rates and differences among combinations

RESEARCH QUESTIONS & EVALUATION

Kyle Williams • Learning to Read Bushman • MSc Research Proposal
RESEARCH QUESTIONS & EVALUATION

• Which segmentation techniques are effective for the machine learning algorithms used in this research?
  • Implement different segmentation techniques
  • Compare accuracy, word error rates and differences among combinations
To what extent can an n-gram language model improve accuracy when automatically transcribing handwritten Bushman texts?

- Attempt to build an n-gram language model
- Compare effect on accuracy, word error rates and differences among combinations
WORK DETAIL

- July 2010: Machine learning algorithms
- October 2010: Segmentation
- 15 September: Background chapter
- January 2011: Language model
- March 2011: Entire system brought together
- May 2011: Final evaluation
- August 2011: Writing of thesis
ANTICIPATED OUTCOME

- Measurement of accuracy for transcription
- Comparison of HMMs, NNs and SVMs and features
- Insight into effective segmentation approaches
- Applicable to other collections
- Software system capable of transcribing Bushman texts
- Transcription of some pages of Bleek and Lloyd Collection
Questions?