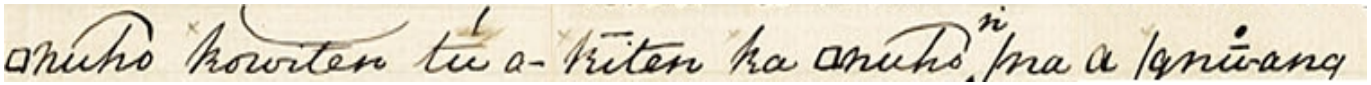
Learning to Read Bushman

Kyle Williams
Supervisor: A/Prof. Hussein Suleman



OUTLINE

- Introduction/Bleek and Lloyd Collection
- Project Description
- Pilot Study
- Methodology
- Practical Implications
- Conclusions

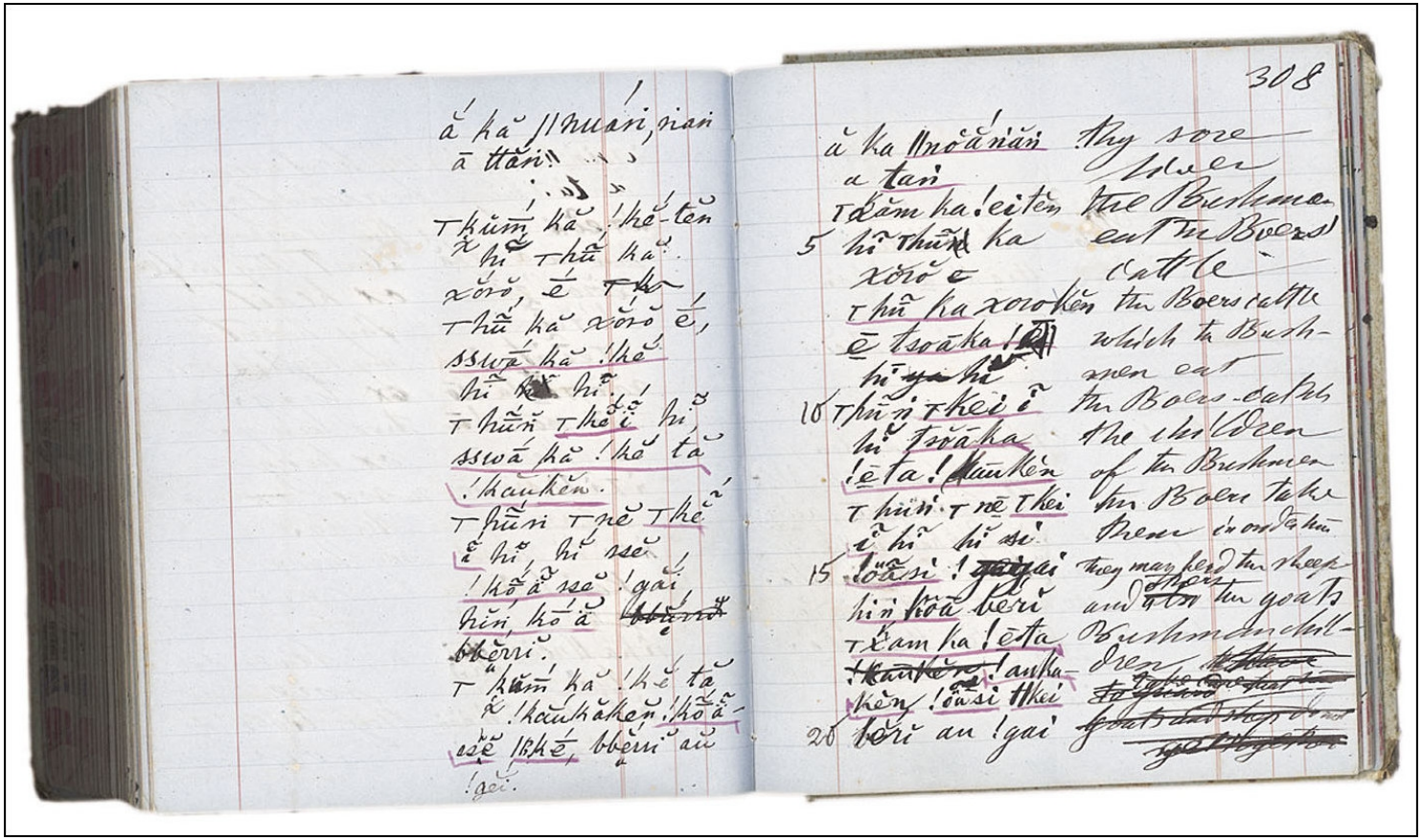
BLEEK AND LLOYD COLLECTION

- 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

BLEEK AND LLOYD COLLECTION



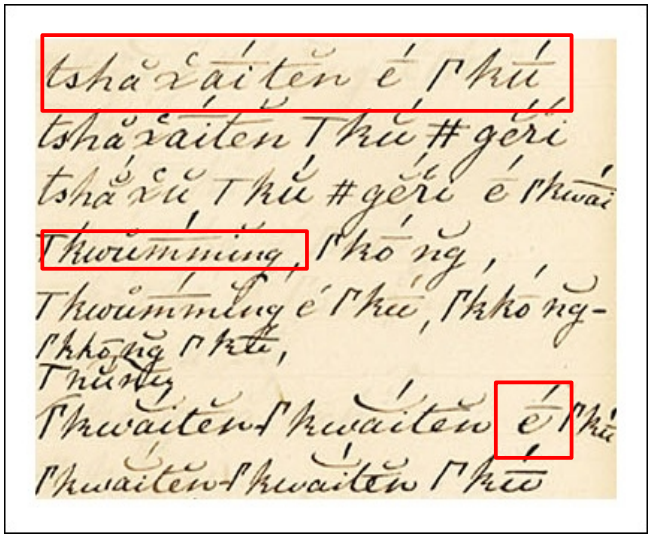
Digital Libraries Laboratory, University of Cape Town

BLEEK AND LLOYD COLLECTION

- Already have systems for preservation and viewing collection
  - Next step involves enhancing use
- Transcription – convert images to text
  - Search, index and compare text
- Transcription is a tedious and time consuming
  - Need for automatic transcription

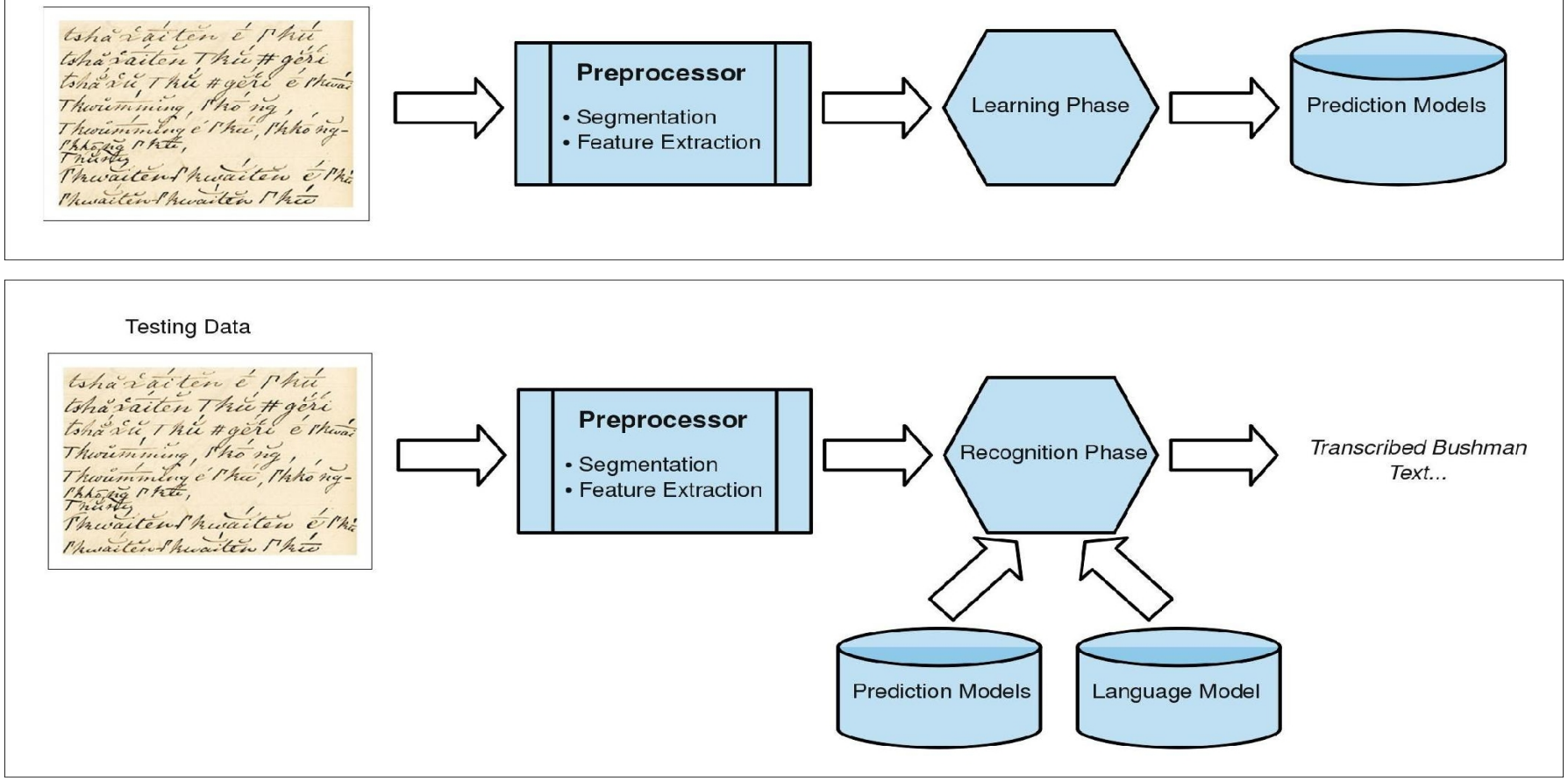
PROJECT DESCRIPTION

- Build a system capable of automatically transcribing handwritten Bushman texts



- Segmentation
- Feature extraction
- Machine learning
  - Hidden Markov Models,
    Neural Networks and Support Vector Machines
- Language Model

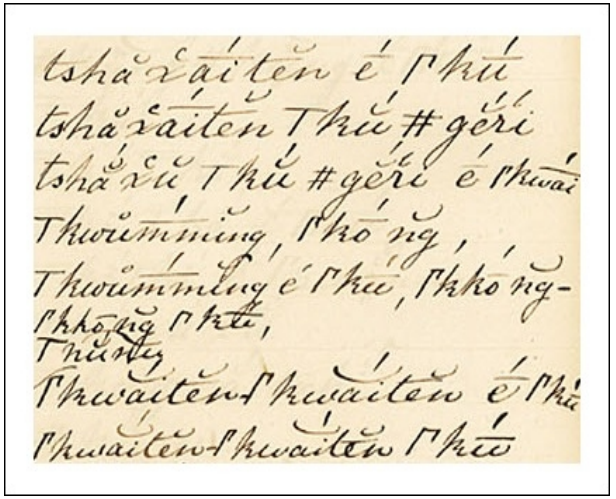
PROJECT DESCRIPTION

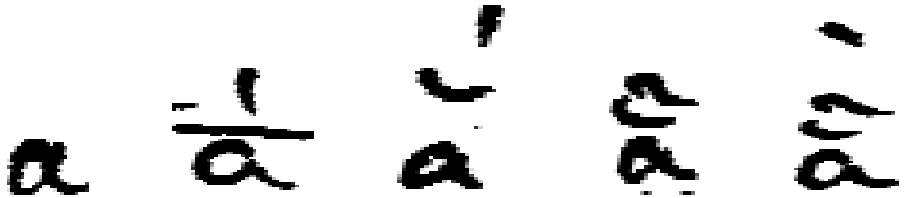


Digital Libraries Laboratory, University of Cape Town

PROBLEM

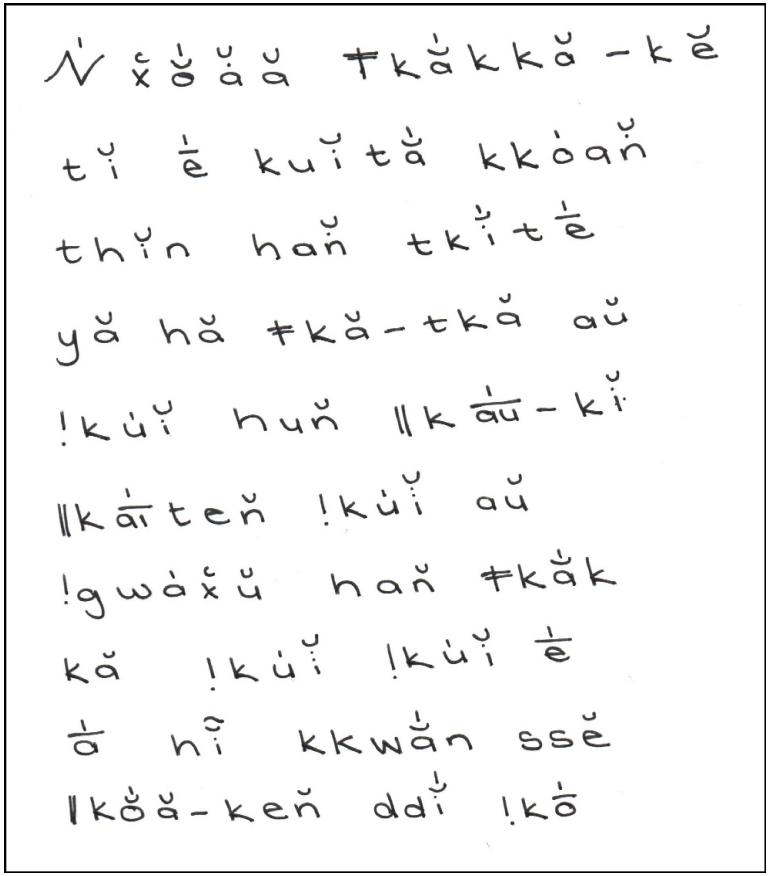
- Diacritics and recognition
  - Multiple diacritics
  - Above and below characters
  - Span multiple characters
- Diacritics and segmentation
- No language model





PILOT STUDY

- Neatly handwritten text
- Limited character set
- Two authors
- 80% transcription accuracy using SVM
- Simplified problem but suggested feasibility



RESEARCH QUESTIONS AND EVALUATION

- How accurately can the automatic transcription of handwritten Bushman texts be performed?
  - Will be answered by addressing 3 sub-questions

RESEARCH QUESTIONS AND EVALUATION

- 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, error rates and differences among combinations

RESEARCH QUESTIONS AND EVALUATION

- Which segmentation techniques are effective for the machine learning algorithms used in this research?
  - Implement different segmentation techniques
  - Compare accuracy, error rates and differences among combinations

RESEARCH QUESTIONS AND EVALUATION

- 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, error rates and differences among combinations, improvements

PRACTICAL IMPLICATIONS?

- Text-to-speech
- Reprinting stories in books
- Automatic translation
- Searchable text
- Insight into Bushman language and culture
- Insight into transcription with diacritics

CONCLUSIONS

- Measurement of accuracy for transcription
- Comparison of HMMs, NNs and SVMs and features
- Insight into effective segmentation approaches
- Insight into language models
- Applicable to other collections

THANK YOU

Questions?

Digital Libraries Laboratory, University of Cape Town