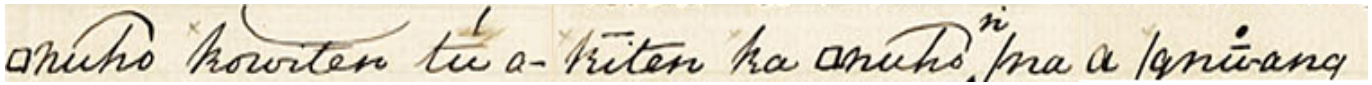
Creating a Handwriting Recognition Corpus for Bushman Languages

Kyle Williams and Hussein Suleman



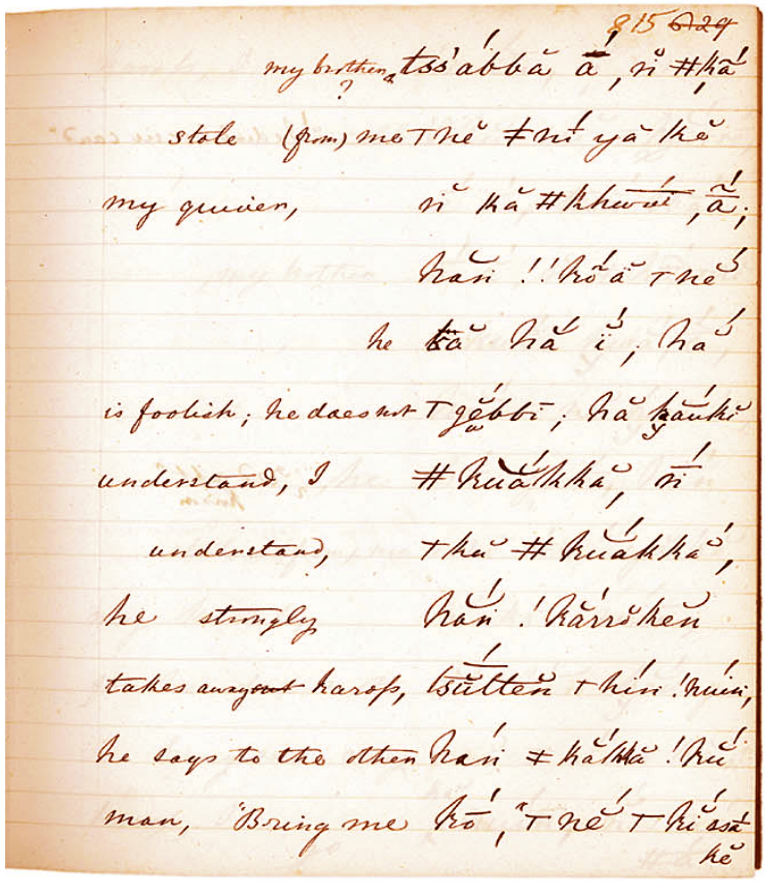
BUSHMAN PEOPLE

- 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



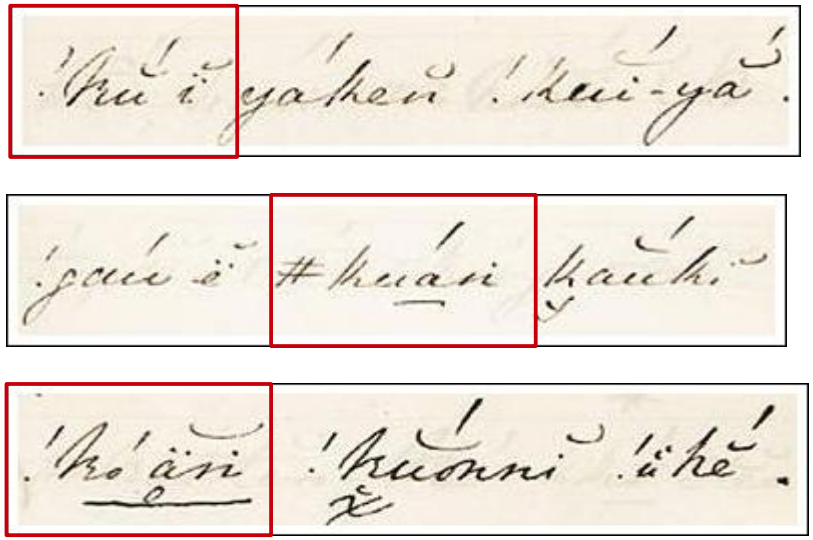
my brothers? tsó'abbă ă, ŭ #hă'
stole (fem) me ǂnĕ ǂnĭ' yă' kĕ
my queen, ŭ kă #khwĭ', ă';
kăn !'kŏ'ă ǂnĕ
he ~~kă~~ kă ĭ; ka
is foolish; he does not ǂgĕbbĭ; kă ǂkaukĭ
understand, I #kuă'kkă, ŭ'
understand, ǂkŭ #kuă'kkă',
he strongly kăn !'kăŭ'kĕn
takes away ~~out~~ hard, tsŭ'ttĕn ǂkĭn !'kuĭn,
he says to the other kăn ǂkăkkă !'kŭ'
man, 'Bring me kŏ', 'ǂnĕ ǂkĭ'ssă'
kĕ

BLEEK AND LLOYD COLLECTION

- Already have systems for preservation and viewing collection
- Next step involves enhancing use
  - Make text searchable
  - Index text
  - Reprint of text in books
  - Text-to-speech
- Need a corpus of transcriptions

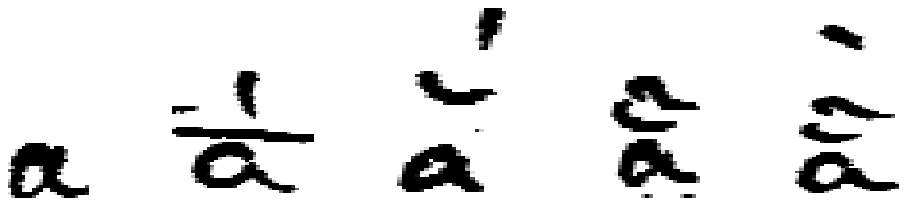
BUSHMAN TEXT

- Text contains complex diacritics
  - Stacked above and below characters
  - Span multiple characters



BUSHMAN TEXT

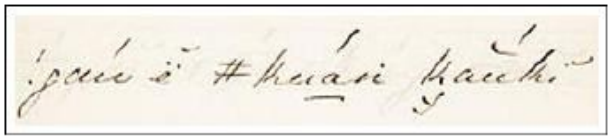
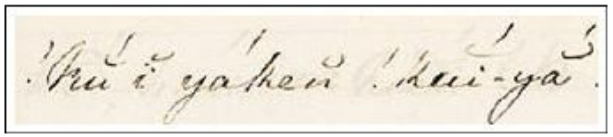
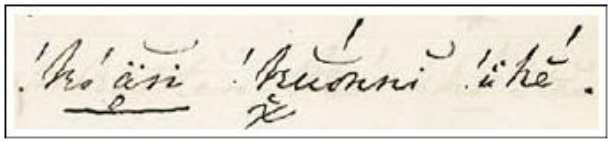
- Diacritics cannot be represented using Unicode
- No one left that speaks the |xam language!
- Over 137 different diacritics (more still being found)



ENCODING

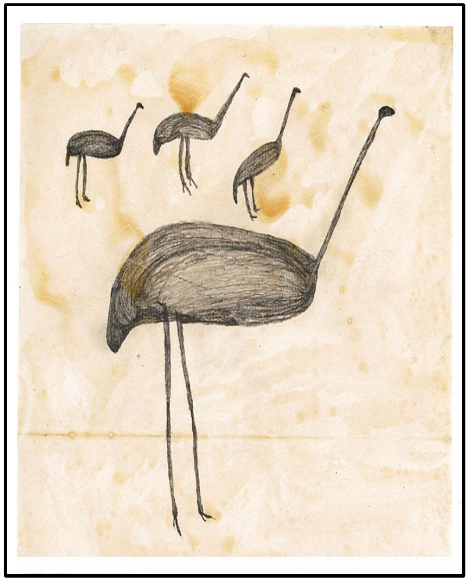
- Bushman text cannot be encoded using Unicode
- Latex IPA package contains diacritics
  - Allows for custom macros to be created
- Stacked, nested, multiple characters
- \uline{a} → $\overset{\shortparallel}{\underset{\smile}{\mathrm{a}}}$
- \xbelow{\uline{a}} → $\underset{\mathrm{x}}{\overset{\shortparallel}{\underset{\smile}{\mathrm{a}}}}$
- \xbelow{aa} → $\underset{\mathrm{x}}{\mathrm{aa}}$

ENCODING

	!ga\dialine{u} \twodotsu{e} \texthash{}ku\barblinet{a}n \ybelow{k}a\uline{u}k\u{i}	!gau̇ ë #kuan̲ k̬aŭkĭ
	!k\uline{u} \twodotsu{i} y\dialine{a}ke\u{n} !ku\uline{i}-y\uline{a} .	!kŭ ï yȧkeň !kui̲-yă .
	!k\barbelow{\dialine{o} \circbtwodotsa{a}\onedot{n}} \xcbelow{k}\uline{uo}nn\u{i} !\uu{u}h\uline{e} .	!kȯäň k̭̰̣̄ ŭonnĭ !ŭhë .

XÒÄ'XÒÄ - “TO WRITE”

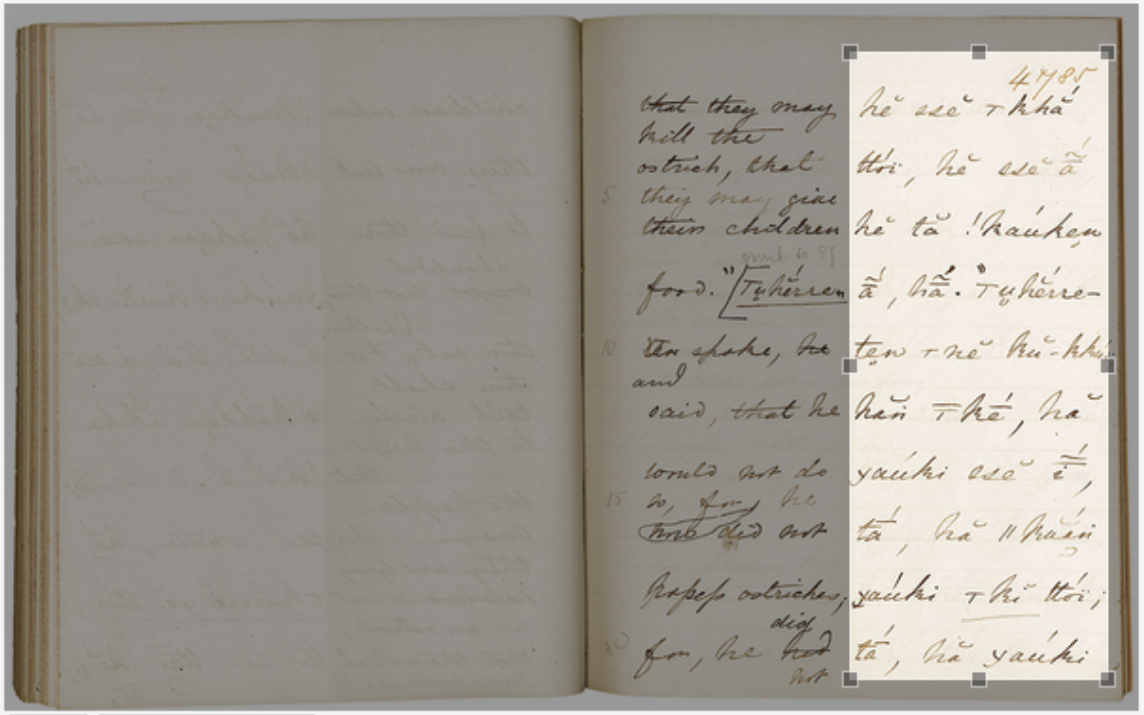
- An AJAX tool to create a Bushman corpus
- Automatic algorithms
- User input
- Preprocessing
- Line and word segmentation
- Transcription
- Job and user management



TEXT SELECTION

Menu

Home
New Segmentation Job
New Transcription Job
Logout



Crop | No Segmentation

Info

Welcome to the segmentation part of the workshop. During this part of the workshop you will assist in checking and correcting the output of automatic page segmentation.

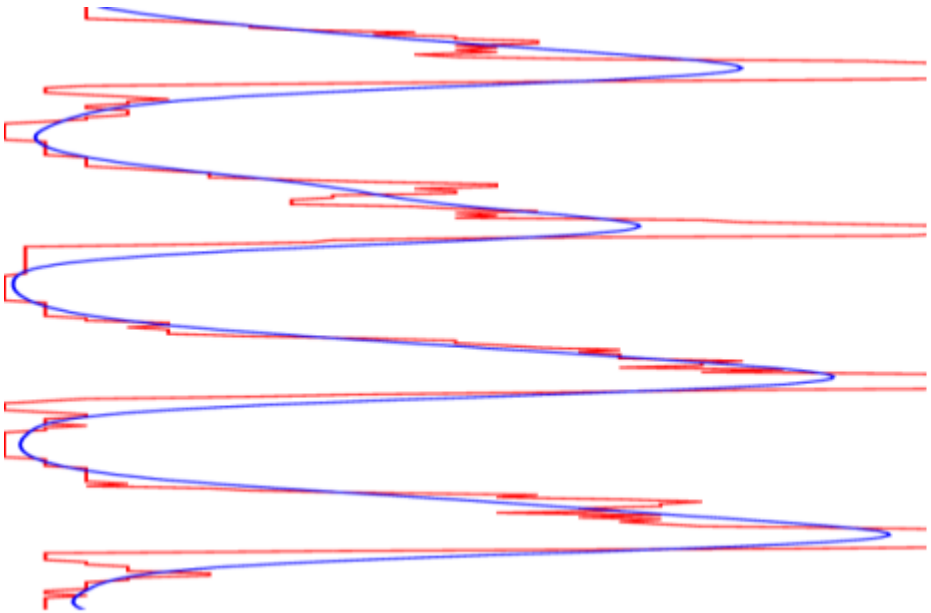
To start, please look at the page on your left, identify the Bushman text on it and draw a rectange around it. Try to get as much Bushman text as possible without getting English text. Don't worry though if you do get English text, we can always fix it later!

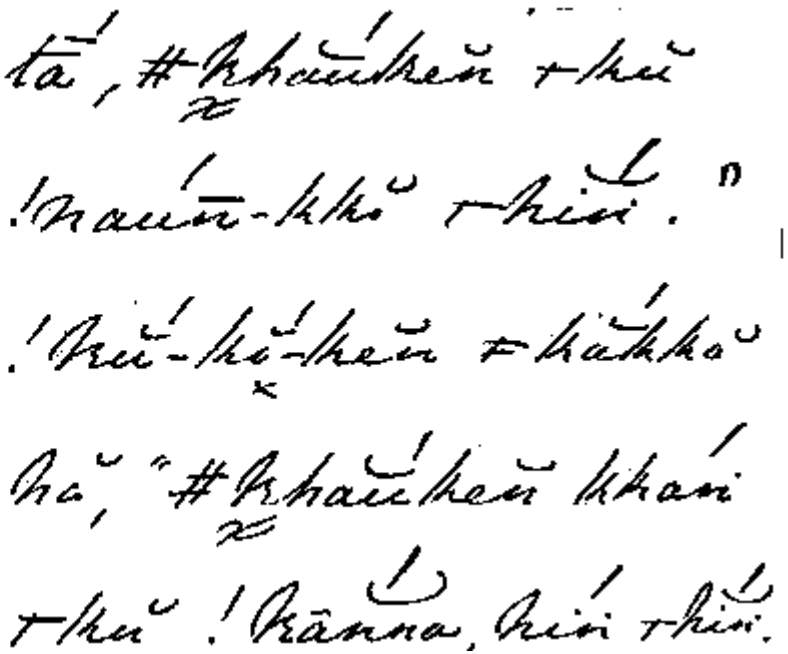
Digital Libraries Laboratory, University of Cape Town

LINE SEGMENTATION

- Projection profile-based line segmentation
- Count foreground-background transitions for each row
- Minima suggest space between lines
  - Could represent space between base character and diacritics
  - Gaussian smoothing of projection profile

LINE SEGMENTATION



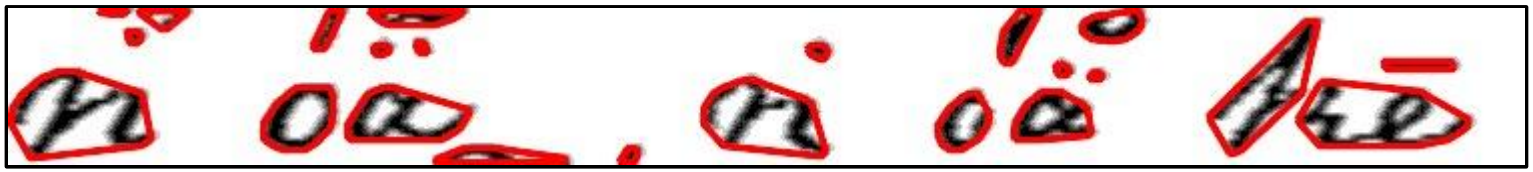


LINE SEGMENTATION

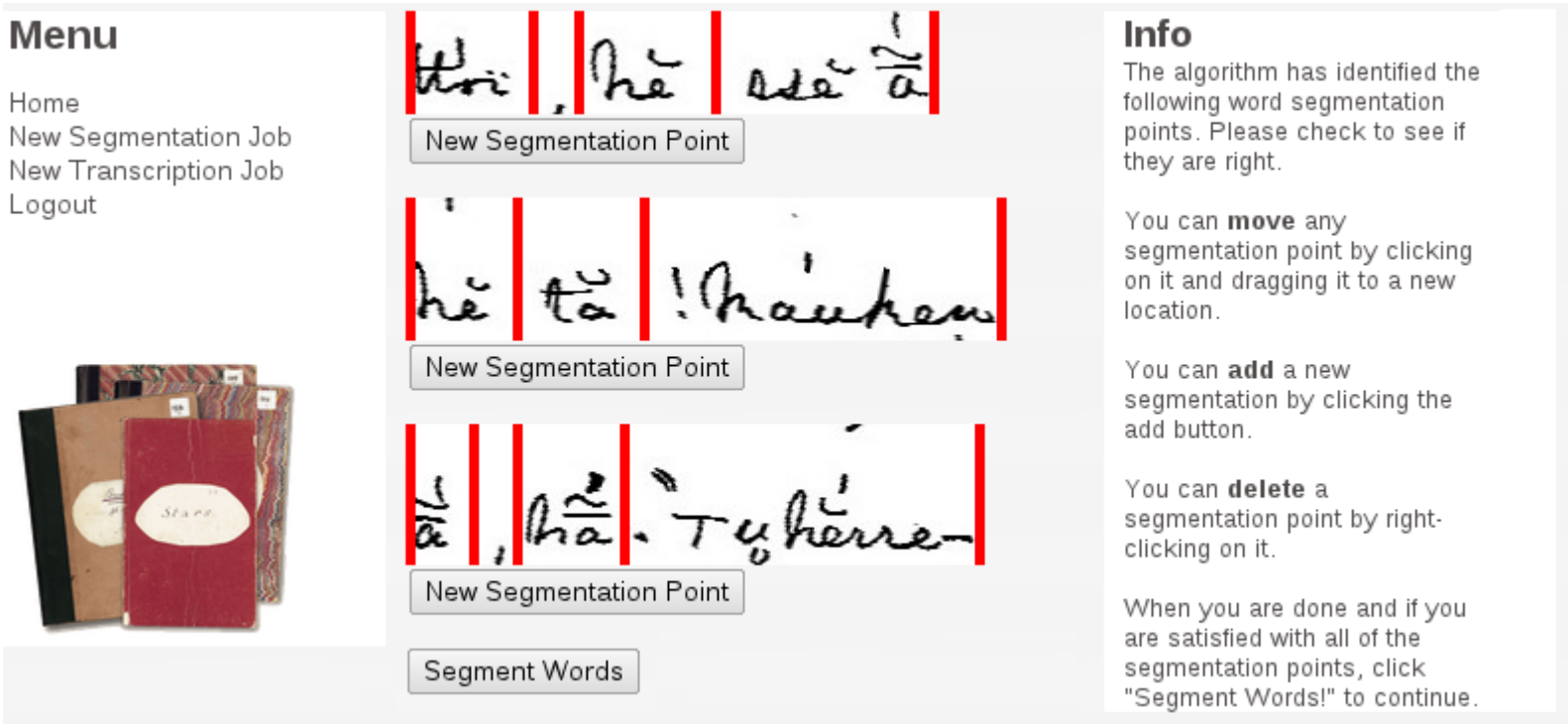
Digital Libraries Laboratory, University of Cape Town

WORD SEGMENTATION

- Line slant is automatically corrected
- Connected components in text lines are identified
- Distances between adjacent components are calculated
- Distances above threshold separate words

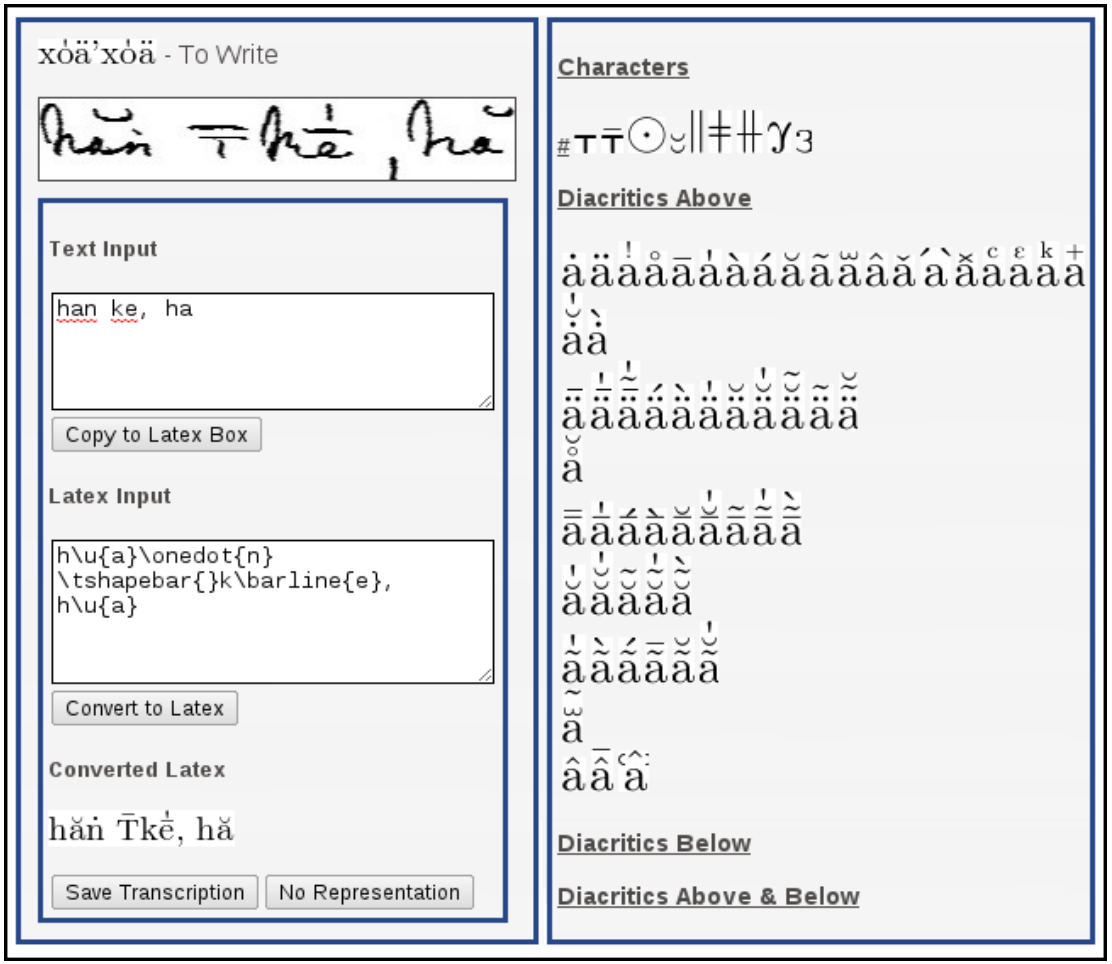


WORD SEGMENTATION



Digital Libraries Laboratory, University of Cape Town

TRANSCRIPTION



Digital Libraries Laboratory, University of Cape Town

CORPUS CREATION WORKSHOPS

- Workshop held to create Bushman corpus
- 29 data capturers recruited
- 900 pages from 2 authors randomly selected
- 729 pages were segmented into lines and words
- 1547 text lines were transcribed
- 452 text lines could not be transcribed
  - Interface didn't support characters, noise, English

CORPUS CREATION WORKSHOPS

- Quality and efficiency of data capturers evaluated
- 5 data capturers asked to return
- 1700 more line recruited
- More efficient and potentially fewer errors

CORPUS CREATION WORKSHOP

Digital Libraries Laboratory, University of Cape Town

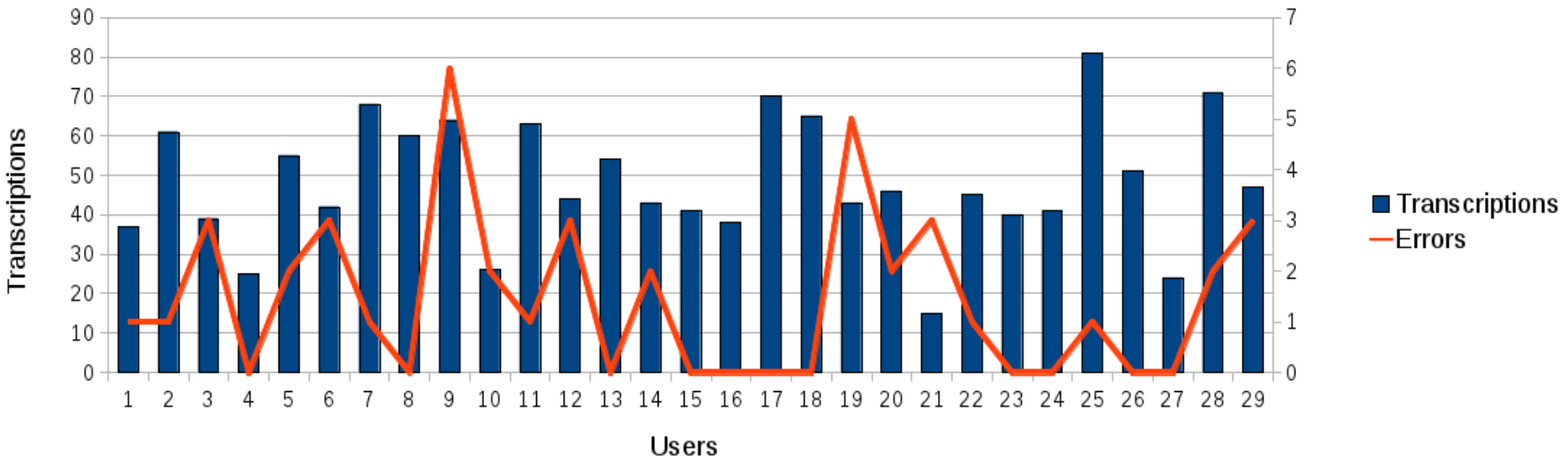
USER CONTRIBUTIONS

Segmentation Jobs		Transcription Jobs	
Jobs	Number of Users	Jobs	Number of Users
0-9	0	0-9	0
10-19	5	10-19	1
20-29	9	20-29	3
30-39	7	30-39	4
40-49	5	40-49	9
50-59	0	50-59	4
60-69	0	60-69	6
70-80	0	70-80	2

DATA QUALITY

- Quality represented by accuracy and correctness of transcriptions
- Useful in planning for follow on workshops
- Random transcriptions by each user evaluated by research assistant
  - Wrong diacritics, characters, etc.
- Average of 0.48 errors per text line
  - Acceptable for lay persons?

EFFICIENCY VS QUALITY



CONCLUSIONS

- Creation of corpora for historical texts is often difficult due to complexities of script
- Semi-automatic tool allowed for more efficient and less expensive creation of corpus
- Currently being used in handwriting recognition study
- Applicable to other historical collections

THANK YOU

Questions?

Digital Libraries Laboratory, University of Cape Town