

# Improving orthographic transcriptions using sentence similarities

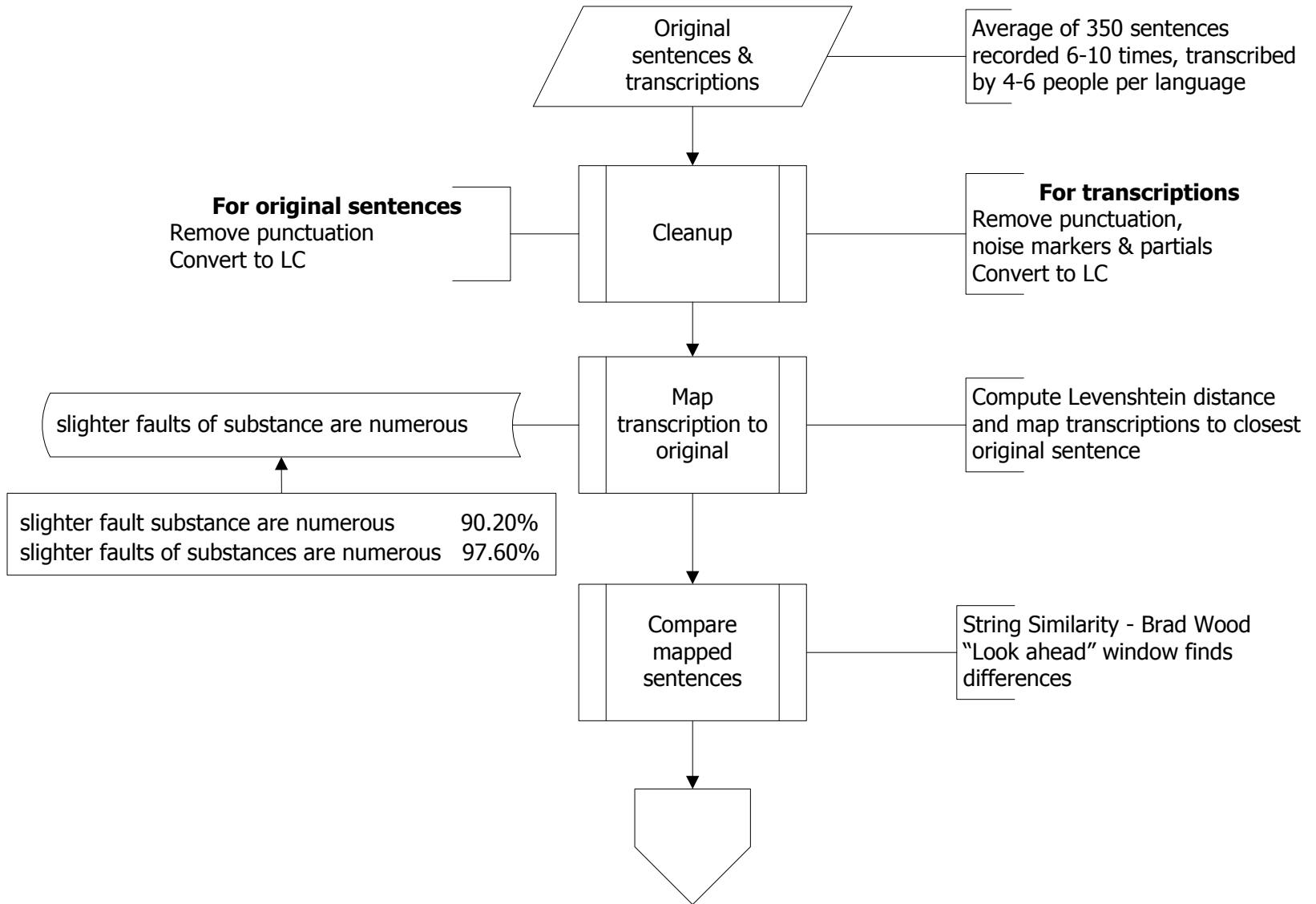
NL Oosthuizen, MJ Puttkammer & M Schlemmer

Centre for Text Technology (CTexT™), North-West University, Potchefstroom Campus (PUK)

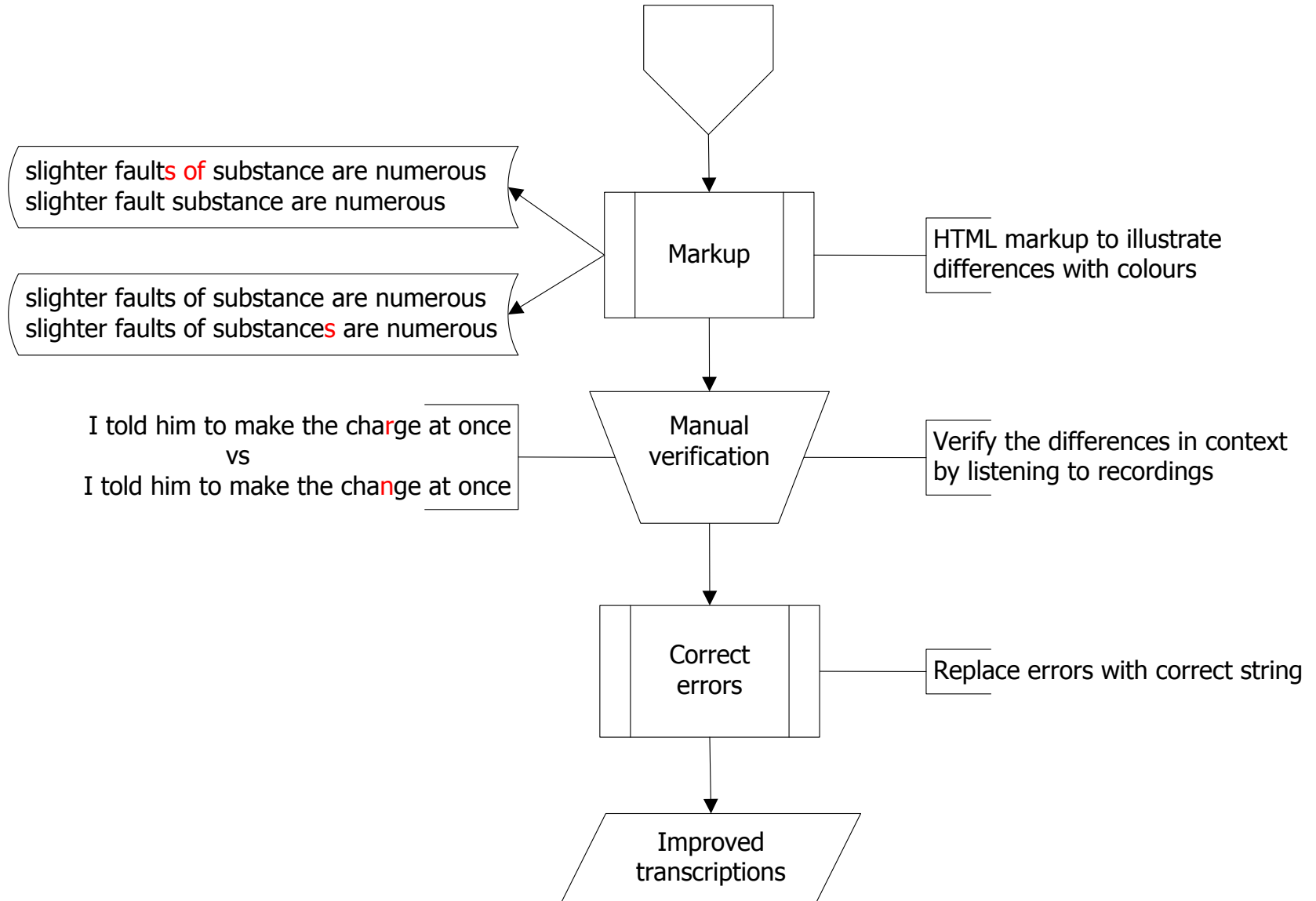
## Introduction

- Multiple transcribers cause inconsistent transcriptions on various levels
  - Confusables
    - has it been tried on too small a scale
    - has it been tried on **to** small a scale
  - Splits
    - there's nowhere else for it to go
    - there's no  where else for it to go
  - Insertions
    - so we took our way toward the palace
    - so we **we\_**took our way toward the palace
  - Deletions
    - as **to\_**the first the answer is simple
    - as the first the answer is simple
  - Non-words
    - there is no **arbitrator** except a legislature fifteen thousand miles off
    - there is no **abritator** except a legislature fifteen thousand miles off
- Ensuring consistency decreases the number of transcription errors

# Process Flowchart I



# Process Flowchart II



# Results

Language	Errors corrected
Afrikaans	152
English	337
isiNdebele	291
isiXhosa	1081
isiZulu	1228
Sepedi	736
Sesotho	261
Setswana	828
Siswati	351
Tshivenda	191
Xitsonga	456