Spellcheckers for Nguni languages

C. Maria Keet

Department of Computer Science
University of Cape Town, South Africa
mkeet@cs.uct.ac.za

MeMaT workshop, 4-5 December 2017

With contributions from

- IsiZulu spellchecker
  - Hussein Suleman
  - Balone Ndaba
  - Langa Khumalo
  - Norman Pilusa
  - Frida Mjaria
- IsiXhosa spellchecker
  - Nthabiseng Mashiane
  - Siseko Neti
  - Norman Pilusa
- Participants in the evaluations from ULPDO@UKZN and Linguistics@UCT
- Additional texts from INC, Mantoa Motinyane-Masoko, MeMaT translators, publicly available texts

Outline

Outline

Motivation

- IsiZulu and isiXhosa, among others, have limited ICTs support
- Most widely spoken languages in South Africa by first language speakers.
- 23% or about 11 million people (isiZulu), 8 million (isiXhosa)

[1] https://www.spellchecker.net/africa_zulu_spell_checker.html

Motivation

- IsiZulu and isiXhosa, among others, have limited ICTs support
- Most widely spoken languages in South Africa by first language speakers.
- 23% or about 11 million people (isiZulu), 8 million (isiXhosa)
- Very limited available spellcheckers for use:
  - Outdated/software not working anymore (Open office plugin)
  - Online one with too many clicks, popups, and ads[1]

[1] https://www.spellchecker.net/africa_zulu_spellchecker.html

General aims

- Investigate development of spellchecker for isiZulu
- Find an approach that can be used across (agglutinating) Bantu languages

What will work best?

- Dictionary approach won't work due to (theoretically) agglutination and (practically) limited dictionaries

What will work best?

- Dictionary approach won't work due to (theoretically) agglutination and (practically) limited dictionaries
- Data-driven statistical model or grammar-based (morphological analyser-based) approach?
- Try that for both isiZulu and isiXhosa

What will work best?

- Dictionary approach won't work due to (theoretically) agglutination and (practically) limited dictionaries
- Data-driven statistical model or grammar-based (morphological analyser-based) approach?
- Try that for both isiZulu and isiXhosa
- Rules for the POS categories coded perform better overall (isiXhosa), but data-driven approach faster and more reusable across languages (isiZulu, isiXhosa), despite being underresourced

Outline

First iteration [Ndaba et al.(2016)]

- Use corpus for data
- Driven by extracted n-gram statistics
- Trigrams and quadrigrams; e.g.
  - *ngimbona*
  - ngi, gim, imb, ...
  - ngim, gimb, ...
- Compute frequencies, determine threshold
- Trigram/quadrigram below threshold: flag words as incorrectly spelled

Basic approach for testing

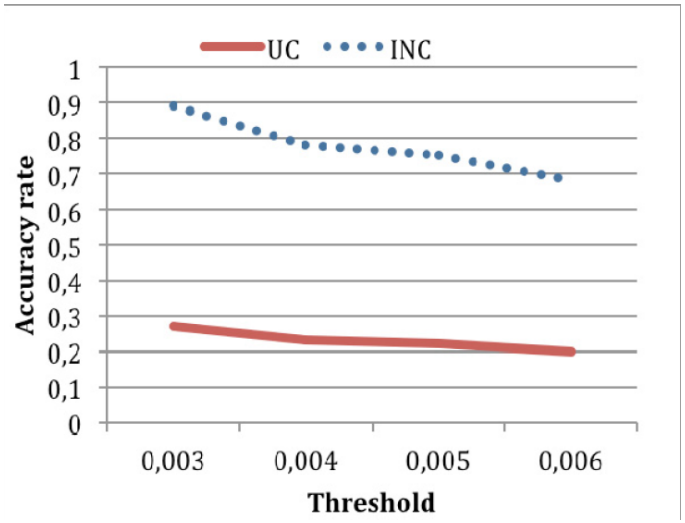
- 10-fold cross-validation for training and testing data set
- 3 corpora to test effect of corpus on accuracy
  - Ukwabelana [Spiegler et al.(2010)]; 288 106 words, 87033 unique
  - Section of the isiZulu National Corpus [Khumalo(2015)]; 538 732 words, 33020 unique
  - IsiZulu news items (collected by MK); 21250 words, 9587 unique
- Different thresholds
- 46 know-to-be-incorrect words added
- Use accuracy as measure, with confusion matrix (TP, TN, FP, FN)

Example

- Corpus trained with:
  - *ngimbona kusasa*: ngi, gim, imb, ...
  - *uvelaphi*: uve, vel, ela, ...
- Spellchecker given following words:
  - *ngivela* $\rightarrow$ accepted, as ngi, vel, and ela are trigrams obtained from the training dataset
  - *Ngivla* $\rightarrow$ rejected, as there is no ivl and no vla

Major outcomes

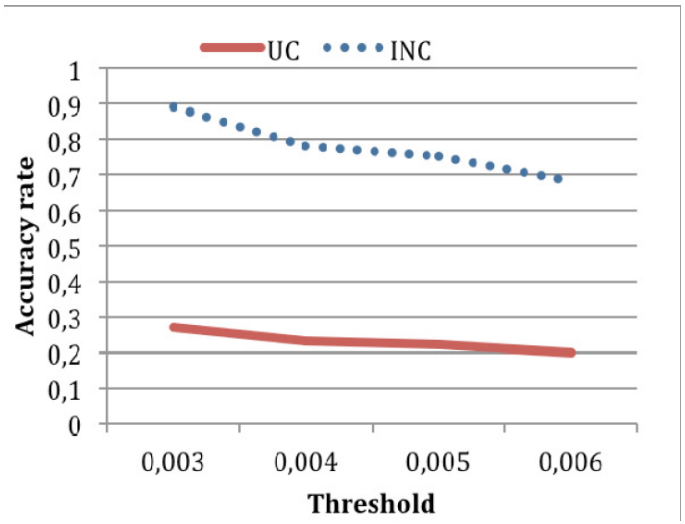
- Accurate in detecting words that do not occur in the training corpus
- The most updated corpora are preferable



- The spellchecker performed slightly better with trigrams than with quadrigrams
- 89% accuracy (on par with older data [Bosch and Eisele(2005), Prinsloo and de Schryver(2004)])

Major outcomes

- Accurate in detecting words that do not occur in the training corpus
- The most updated corpora are preferable



- The spellchecker performed slightly better with trigrams than with quadrigrams
- 89% accuracy (on par with older data [Bosch and Eisele(2005), Prinsloo and de Schryver(2004)])
- Tests with more data: 83% accuracy with noisy data, 85% with cleaned data

Try this with isiXhosa

- Use code for isiZulu, but feed it isiXhosa texts to create a language model for isiXhosa
- Determine threshold
- Determine accuracy

Try this with isiXhosa: results

- 20K tokens corpus, mainly medical documents
- Threshold: 0.002 (marginally better than 0.003)
- Accuracy: about 79%
- (current implemented version trained with much more text)

Error correction for isiZulu

- Statistical language model based approach as well

Error correction for isiZulu

- Statistical language model based approach as well
- Insertions, deletions, transpositions, substitutions

Error correction for isiZulu

- Statistical language model based approach as well
- Insertions, deletions, transpositions, substitutions
- Levenshtein distance + probability of alternate trigram

- Probabilities of successive trigrams

Error correction for isiZulu

- Statistical language model based approach as well
- Insertions, deletions, transpositions, substitutions
- Levenshtein distance + probability of alternate trigram
  - nii
  - ngi
  - distance: 1 (a substitution of i where there should be g)
- Probabilities of successive trigrams

Error correction for isiZulu

- Statistical language model based approach as well
- Insertions, deletions, transpositions, substitutions
- Levenshtein distance + probability of alternate trigram
  - nii
  - ngi
  - distance: 1 (a substitution of i where there should be g)
- Probabilities of successive trigrams
- Measures:
  - Can it propose something (C_s)?
  - Is the intended word among the proposed words (C_v)? (relevance)

Results

- It can propose something quite well for each type of typo ($<$ 90% accuracy)
- The relevance varies a lot:
  - Substitutions 59%
  - Insertions 30%
  - Deletions 73%
  - Transpositions 89%

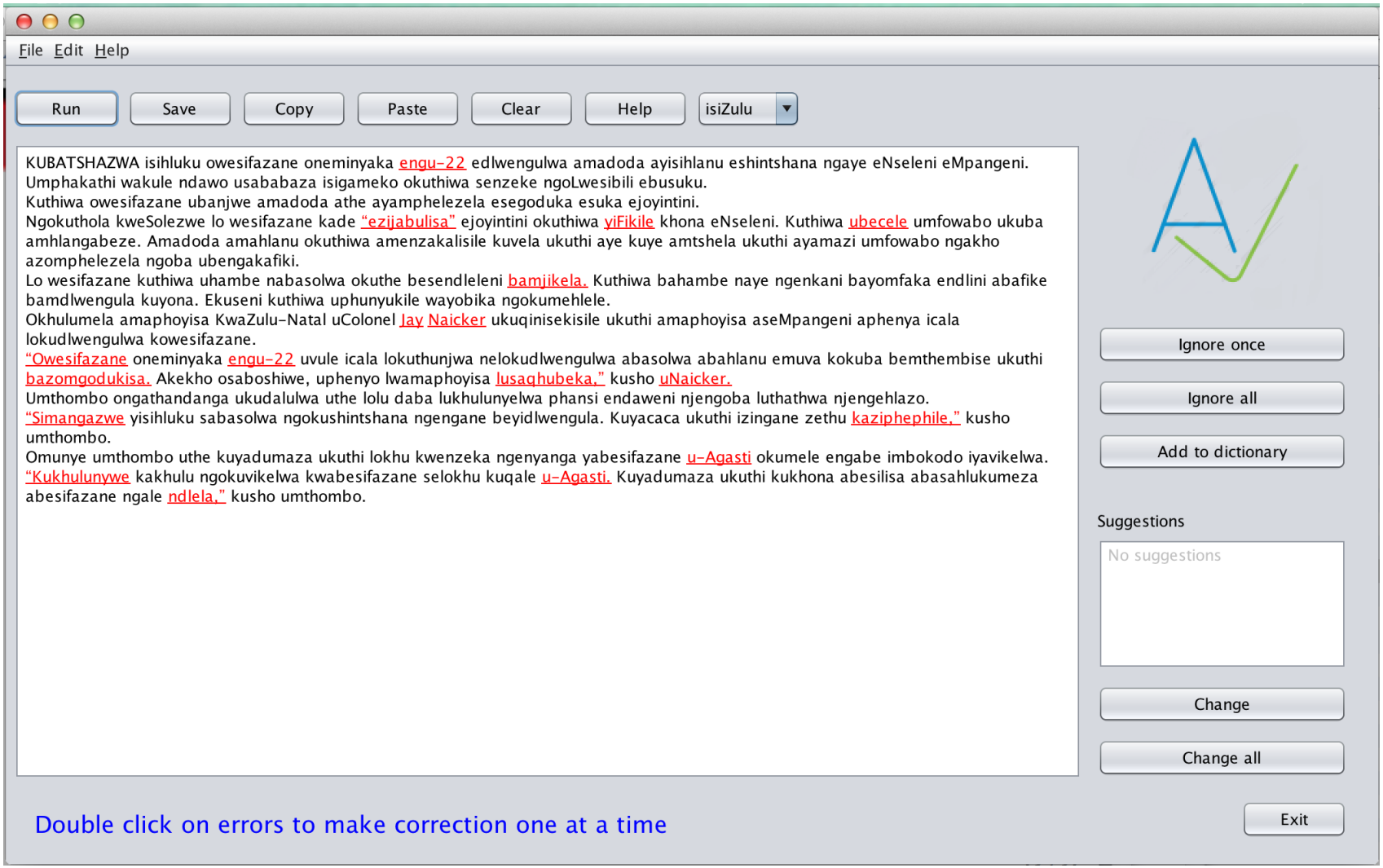
Results

- It can propose something quite well for each type of typo ($<$ 90% accuracy)
- The relevance varies a lot:
  - Substitutions 59%
  - Insertions 30%
  - Deletions 73%
  - Transpositions 89%
- Why? We don't know for sure yet

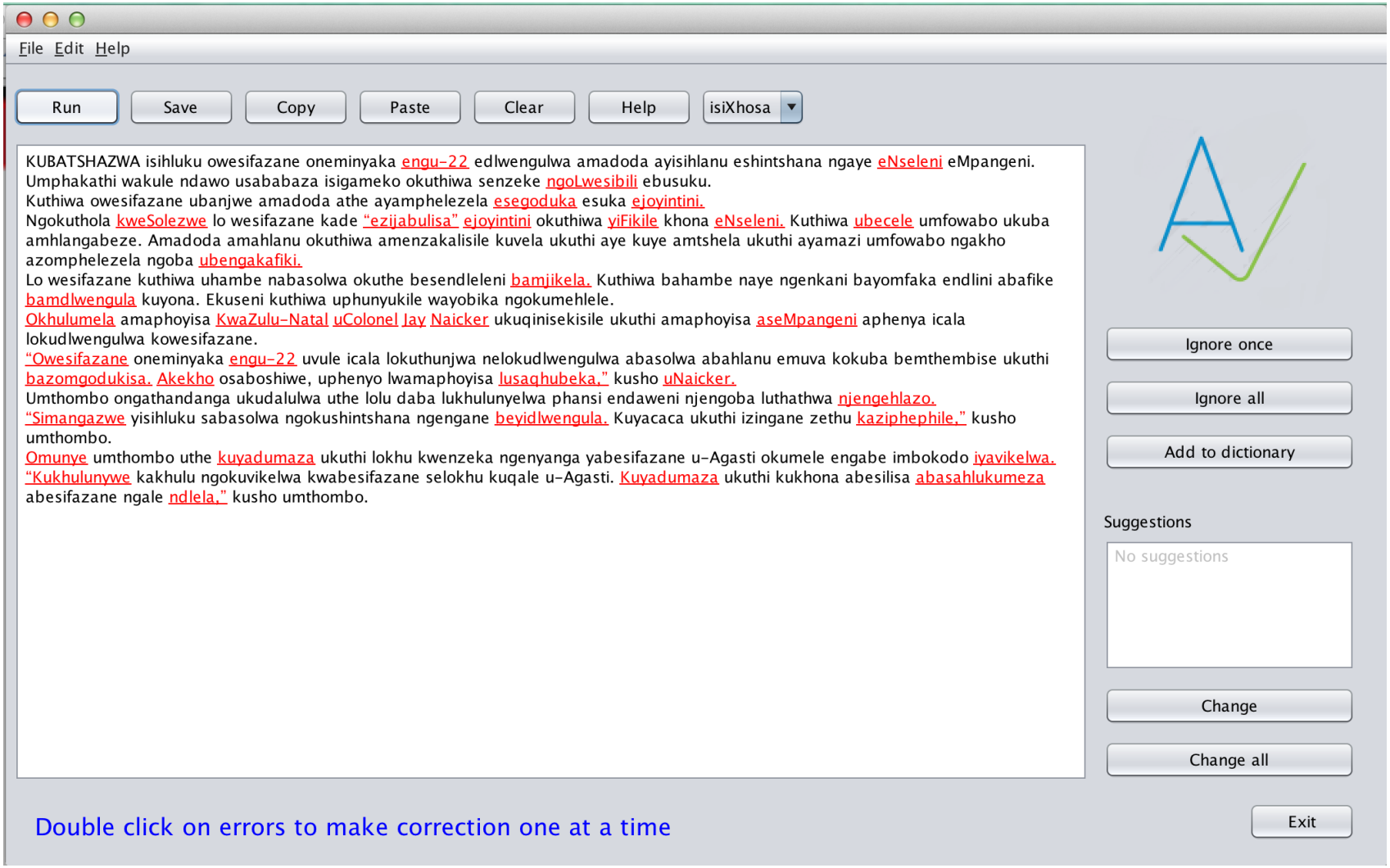
Spelling correction for isiXhosa

- Used the same code as for isiZulu
- But with the isiXhosa language model
- Implemented, but no idea on effectiveness yet

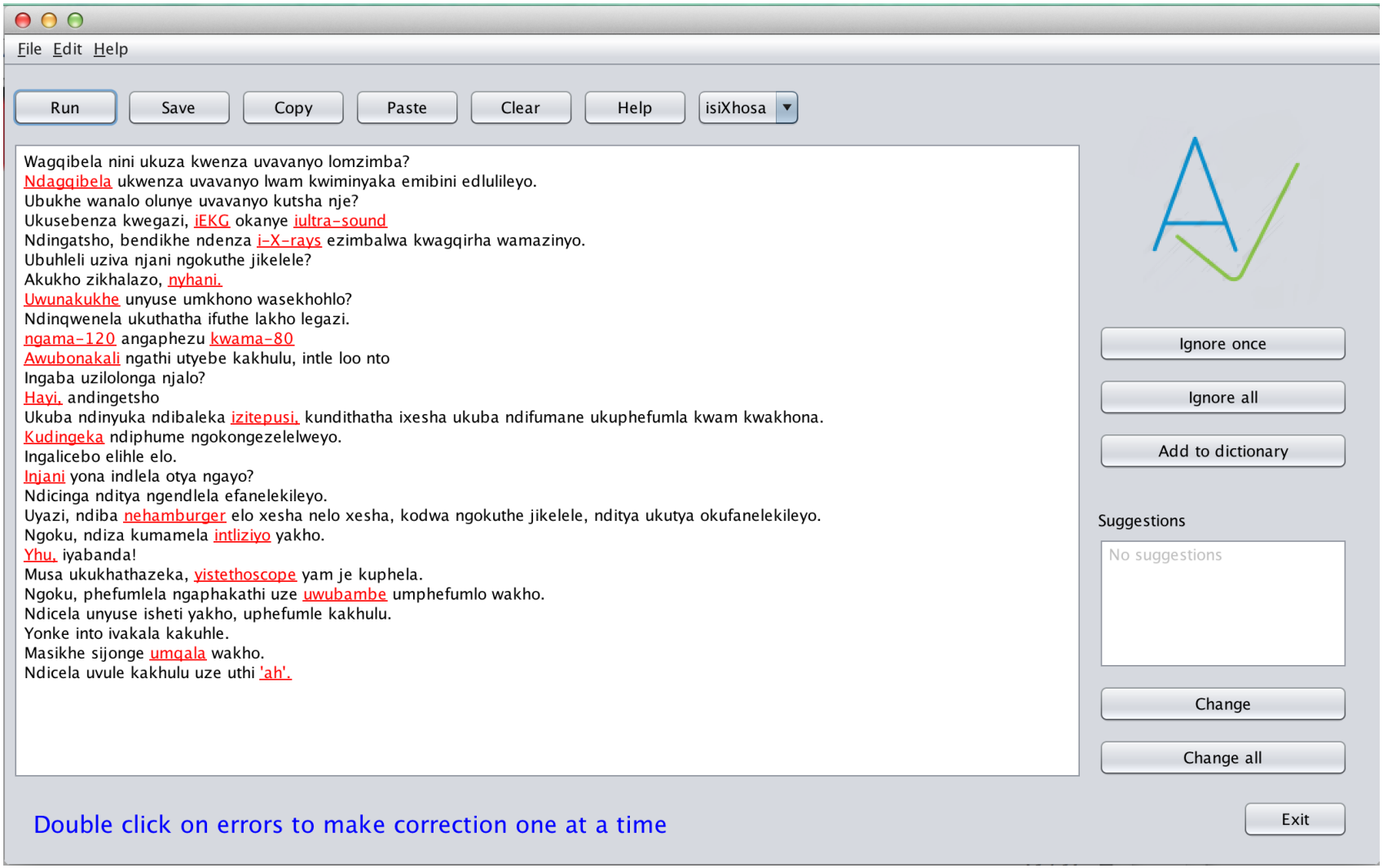
isiZulu text, isiZulu model



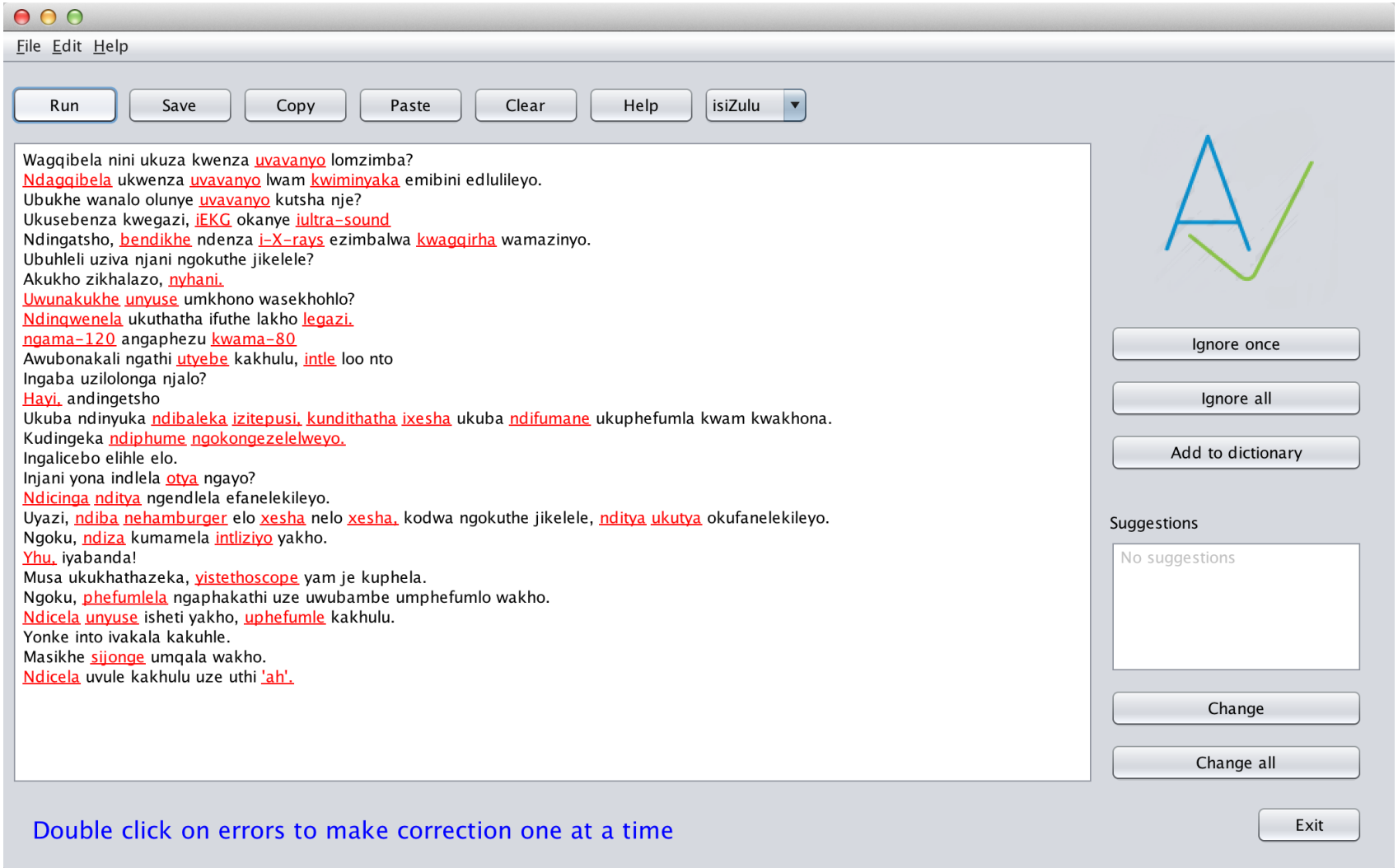
isiZulu text, isiXhosa model



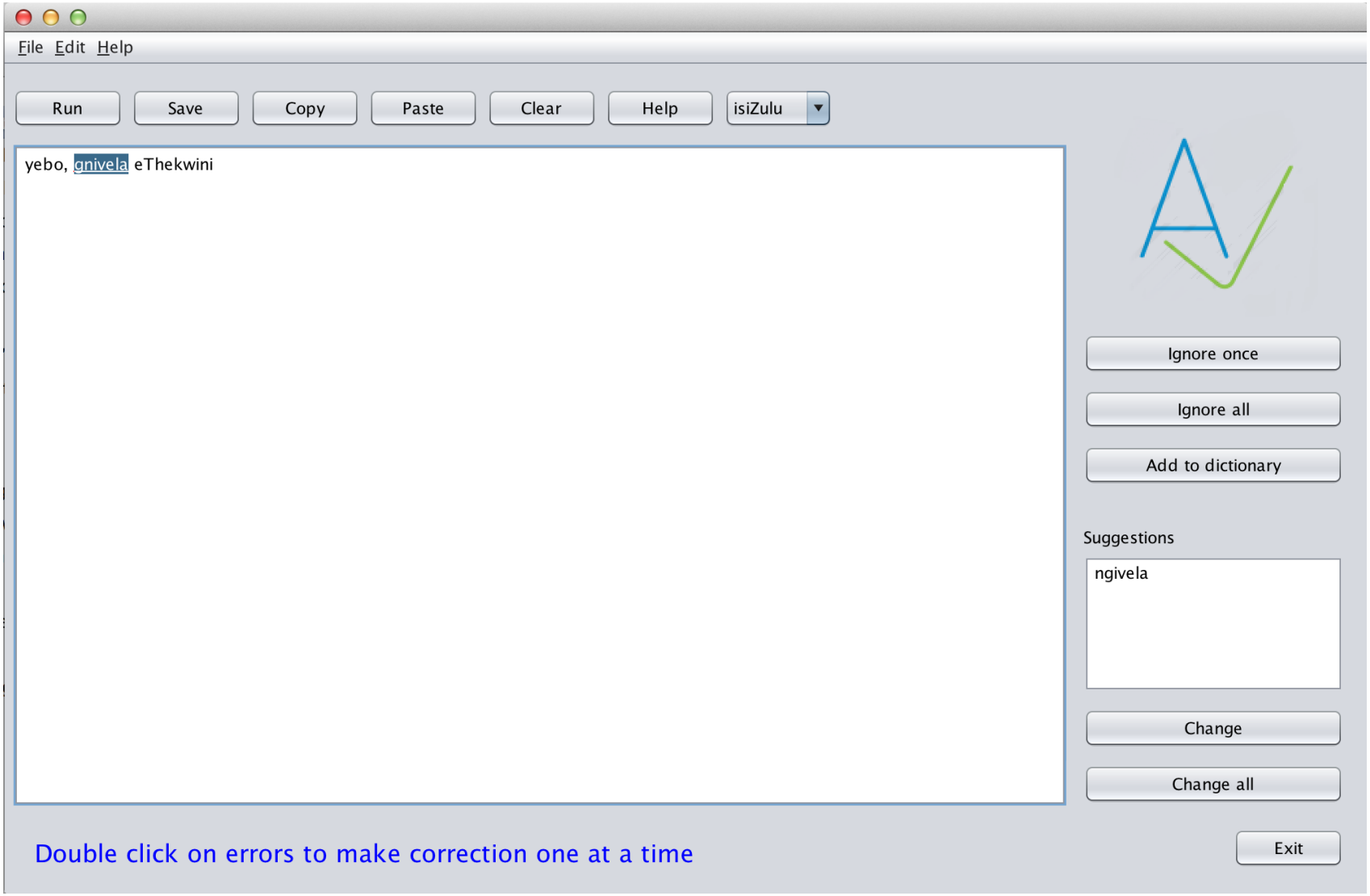
isiXhosa text, isiXhosa model



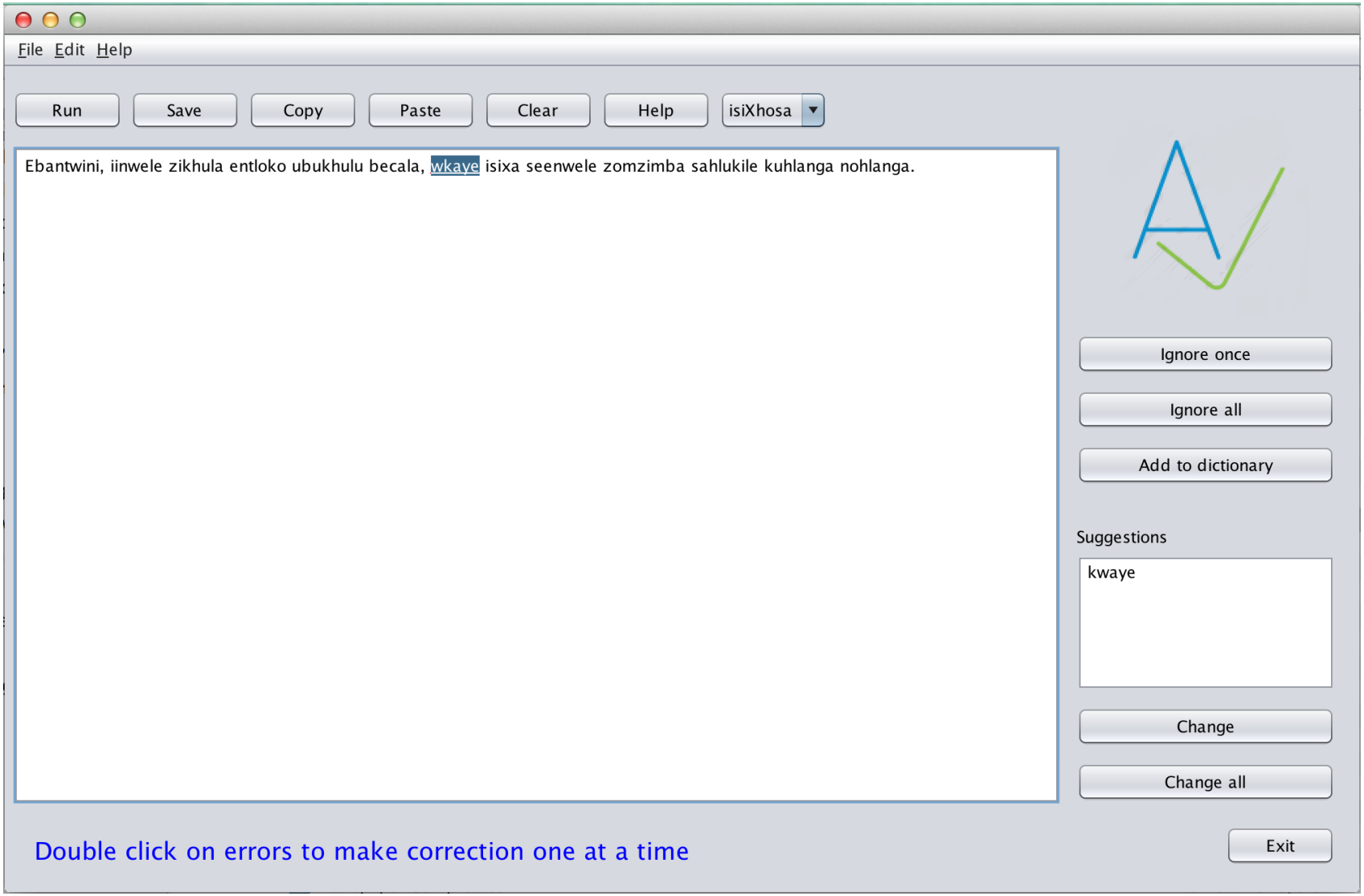
isiXhosa text, isiZulu model



Error correction–transposition typo



Error correction–transposition typo



Outline

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?
- Corpus size? Genre?
- Timeliness of the text

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?
- Corpus size? Genre?
- Timeliness of the text
- Lowercase vs upper case (e.g., *eGoli*)

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?
- Corpus size? Genre?
- Timeliness of the text
- Lowercase vs upper case (e.g., *eGoli*)
- Sociolinguistics, if dialects have words written differently

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?
- Corpus size? Genre?
- Timeliness of the text
- Lowercase vs upper case (e.g., *eGoli*)
- Sociolinguistics, if dialects have words written differently
- Room for improvement on the corrector

Discussion

- Setting the threshold is difficult
  - Cleaned data or noisy data?
  - Trigrams on text proper or on non-punctuation-marks?
  - Cleaned trigrams or not?
- Corpus size? Genre?
- Timeliness of the text
- Lowercase vs upper case (e.g., *eGoli*)
- Sociolinguistics, if dialects have words written differently
- Room for improvement on the corrector
- How much do the isiZulu and isiXhosa language models differ?

Outline

Conclusions

- We now have the spellcheckers
- Reasonable accuracy
- Outstanding questions on generalisability and improvement of accuracy
- Would it enhance MT or introduce more noise?

References I

Sonja E. Bosch and Roald Eisele.

The effectiveness of morphological rules for an isiZulu spelling checker.
South African Journal of African Languages, 25(1):25–36, 2005.

Langa Khumalo.

Advances in developing corpora in African languages.
Kuwala, 1(2):21–30, 2015.

B. Ndaba, H. Suleman, C. M. Keet, and L. Khumalo.

The effects of a corpus on isizulu spellcheckers based on n-grams.
In Paul Cunningham and Miriam Cunningham, editors, *IST-Africa 2016*. IIMC International Information Management Corporation, 2016.
11-13 May, 2016, Durban, South Africa.

D. J. Prinsloo and G.-M. de Schryver.

Spellcheckers for the south african languages, part 2: the utilisation of clusters of circumfixes.
South African Journal of African Languages, :83–94, 2004.

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach.

Ukwabelana – an open-source morphological zulu corpus.
In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 1020–1028. Association for Computational Linguistics, 2010.
Beijing.

Thank you!

Questions?