1/43

# 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



- 2 Data-driven spellcheckers
  - Error detection for isiZulu
  - Error detection for isiXhosa
  - Error correction for isiZulu and isiXhosa
  - Demo

#### Oiscussion



## Outline



- 2 Data-driven spellcheckers
  - Error detection for isiZulu
  - Error detection for isiXhosa
  - Error correction for isiZulu and isiXhosa
  - Oemo

#### 3 Discussion

#### 4 Conclusions

## 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)

<sup>&</sup>lt;sup>1</sup>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  - Outdated/software not working anymore (Open office plugin)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<sup>1</sup>

<sup>1</sup>https://www.spellchecker.net/africa\_zulu\_spell\_checker.html = \_\_\_\_\_\_

(日) (四) (三) (三) (三)

7/43

#### General aims

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

(日) (四) (三) (三) (三)

8/43

# 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

## Motivation

- 2 Data-driven spellcheckers
  - Error detection for isiZulu
  - Error detection for isiXhosa
  - Error correction for isiZulu and isiXhosa
  - Demo

#### 3 Discussion

#### 4 Conclusions

イロト 不得 とくき とくきとう き

12/43

# 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)

14/43



- Corpus trained with:
  - ngimbona kusasa: ngi, gim, imb, ...
  - uvelaphi: uve, vel, ela, ...
- Spellchecker given following words:
  - $\textit{ngivela} \rightarrow \text{accepted},$  as ngi, vel, and ela are trigrams obtained from the training dataset
  - $\mathit{Ngivla} 
    ightarrow$  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

17/43

# 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



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

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

• Probabilities of successive trigrams

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

- Statistical language model based approach as well
- Insertions, deletions, transpositions, substitutions
- Levenshtein distance + probability of alternate trigram
  - nii
  - ngi
  - $\bullet\,$  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<sub>v</sub>)? (relevance)

イロト 不得 とくき とくきとう き

24 / 43

## Results

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

イロト 不得 とくき とくき とうき

25/43

## Results

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

Conclusions

## 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

000	
<u>E</u> lle <u>E</u> dit <u>H</u> elp	
Run         Save         Copy         Paste         Clear         Help         IsiZulu           KUBATSHAZWA isihuku owasifazane one minyaka engu-22 diwenguka amadoda ayisihanu eshinshana ngaye ekiseleni eMpangeni.         Umphakathi wukule nakawo usiabhaza isiganeko kokuthiwa senzeke ngotwesibili ebusuku.           KUBATSHAZWA isihuku onakon usiabhaza isiganeko kokuthiwa senzeke ngotwesibili ebusuku.         Kuthwa owesifazane onakaha kade "zigabhada" esepaduka esuka ejovinin.           Myakutha lawo subabaza isiganeko kokuthwa senzeke ngotwesibili ebusuku.         Kuthwa owesifazane onakahada kade "zigabhada" kuwala ukuth aye koya antshela ukuthi ayamazi umfowaho ukuba azompheizela ngoba ukupangaka kuthwa amenzakalisile kuwala ukuthi aye koya antshela ukuthi ayamazi umfowaho agakho azompheizela ngoba ukupangaka kuthi akaka kuwala ukuthi aya kuya antshela ukuthi ayamazi umfowaho akuba bada wa menzakalisile kuwala ukuthi aya kuya antshela ukuthi ayamazi umfowaho makho azompheizela ngoba ukutha abadowa kuthe beendleeni gamileka. Kuthwa tamben naye ngenkani bayomfaka endlini abafike Lowesi akutha ngahoyisa aseMazangeni aphenya izala lokudhwenguku abasohwa kuthe shakatare.           Obusuhengua hanphoyis KazaZutha. Nali ukuchen Jikakati kukui hanaphoyisa aseMazane.         Nali ukutha hankatare ngahangeni aphenya izala           Obusuhengua kuyona. Ekuseni kuthiwa uphukuiku kukatar ukui hikuho hakatare.         Nali ukutha hankatare ngahangeni aphenya izala           Obusuhengua kuyona kukuni kuku kukutha ngakunbeka.         Nali ukutha hankatare.         Nali ukutha hankatare.           Obusuhengua kuyona kukuni kuku kukutha ngakunbeka.         Nali asasohakukutha tampanoyisa hasahankatar	Ignore once Ignore all Add to dictionary Suggestions No suggestions Change all
Double click on errors to make correction one at a time	Exit

## isiZulu text, isiXhosa model

000	
<u>Eile E</u> dit <u>H</u> elp	
Run         Save         Copy         Paste         Clear         Help         isXhoa           KURATSHAZWA isihuku owesifazane oneminyaka gngi-22, ediwengukwa amadoda ayisihanu eshintshana ngaye eXiseleni eMpangeni.         Umphakathi wakule ndawo usababaza isigameko kokultiwa senzeke ngakwesihii ebosuku           Kutart SHAZWA isihuku owesifazane oneminyaka gngi-22, ediwengukwa amadoda ayisihanu eshintshana ngaye eXiseleni eMpangeni.         Umphakathi wakule ndawo usababaza isigameko kokultiwa senzeke ngakwesihii ebosuku           Kutart SHAZWA isihuku omesifazane undoda ati waymphetzeda gstopdial esuka ngotimet.         Kutiwa owesifazane undoka ati wa amenakalisile kukula usutima yaymazi unfowako ngaku           Lowesifazane kandodi amahabatoka kukuha amenakalisile kukula isaymazi unfowako ngaku         Sucompakatika undoka ati kando kanda amenakalisile kukula isaymazi unfowako ngaku           Dawesifazane kukuna utartiwa undoka atabatoka kukuha ngokumehide.         Ostati angato meninyaka sukula kukula kukula ngokumehide.           Ostati angato meninyaka sukula kukula kukula ngokumehide.         Ostati angato meninyaka sukula kukula kukalike kukula ngokumehide.           Ostati angato meninyaka sukula kukula kukalike kukula ngokumehide.         Ostati angato meninyaka sukula kukula kukici ukukula ngokumehide.           Ostati angato meninyaka sukula kukula kukici ukukula ngokumehide.         Ostati angato meninyaka sukula kukula kukula kukula kukula kukula kukula kukula kukula.           Ostati angato meninyaka sukula kukula kukula kukula ngaka petih kukula kukula ngange telakukukuter.         Ostati angato meninyaka su	Ignore once Ignore all Add to dictionary Suggestions No suggestions Change all
Double click on errors to make correction one at a time	Exit

#### isiXhosa text, isiXhosa model

00	
Eile Edit Help	
Run       Save       Copy       Paste       Clear       Help       IstXhosa         Wagqibela nini ukuza kwenza uwavnyo komzimba?         Magqibela ukwenza uwavnyo kam kwiminyaka emibni edlulleyo.         Ubushe wando kuminya kuma kuma kuma kuma kuma kuma kuma kum	Ignore once Ignore all
<u>Kudingeka</u> ndiphume ngokongezelelweyo. Ingalcebo elihe elo. Ingan yona nidela obya ngayo? Ndicinga ndiya ngendlela efanekkileyo. Uwaz. ndiha nehamburger elo xesha nelo xesha, kodwa ngokuthe iikelele, nditya ukutya okufanelekileyo.	Add to dictionary
Vigetu, ni dra kalandaraki, ko kusin kusina, kusina ngokulie jiketer, nunya kusiya okulanetekneyo. Yuu yabanda Musa kukuhathazeka, <u>wistefinoscone</u> yan je kuphela. Ngoku, phetimita ngophakatin uze <u>wukatambe</u> umphefumio wakho. Ndicela unvus kala kakuhe. Masike sijonge <u>umgal</u> a wakho.	No suggestions
	Change Change all
Double click on errors to make correction one at a time	Exit

Conclusions

# isiXhosa text, isiZulu model

Elle Edit Help	
Run Save Copy Paste Clear Help isiZulu 🔻	
Wagabela inii ukuza kwenza uowango lonzimba? Magabela inii ukuza kwenza uowango kusha nje? Uluukhe wanalo olume uowango kusha nje? Uluukhe wanalo olume uowango kusha nje? Uluukhe uzwa njesa i, j£KC dame wa lukta-somd Walingatabo, bandikhe ndena i.X-rays ezimbalwa kwapatha wamazinyo. Uluuhei uzwa njeni ngokuthe jikele? Akuboa sha uno mbala. Manawena kukhatha infune lako legazi. ngama-120 angaphezu kwama-80 Awubonakii ngokuthe jikele? Manawena kukhatha infune lako legazi. Ingatabo elimiyata ndibaleka izitepusi. kundithatha izesha ukuba ndifumane ukuphefumia kwam kwakhona. Ukuba ndinyaka ndibaleka izitepusi. kundithatha izesha ukuba ndifumane ukuphefumia kwam kwakhona. Ukuba ndinyaka ndibaleka izitepusi. kundithatha izesha ukuba ndifumane ukuphefumia kwam kwakhona. Ukuba ndinyaka ndibaleka izitepusi. kundithatha izesha ukuba ndifumane ukuphefumia kwam kwakhona. Ukuba ndinyaka ndibaleka izitepusi. kundithatha izesha ukuba ndifumane ukuphefumia kwam kwakhona. Unaya indiba oba ngayo? Malicinga ndiha ngamela ha nganela ka nda nga ukupa okufa nelekileyo. Nyazi, ndiha nehamburger elo xasha, nelo waha, kodwa ngokuthe jikelele, <u>ndinya ukutya</u> okufanelekileyo. Nyaku, ndiha numaki initizya vaho. Tina, yabanda! Majaku phathati ngaphakathu ize uwbambe umphefumio wakho. Ndicka nunga kukuhe. Masikhe sijonge umgala wakho. Ndicka nunga umgala wakho.	Ignore once Ignore all Add to dictionary Suggestions No suggestions Change
Double click on errors to make correction one at a time	Exit

#### Error correction-transposition typo

Go, <u>autovan</u> e i nekwini	A/
	Ignore once
	Add to dictionary
	Suggestions
	ngivela
	Change
	Change all

31/43

#### Error correction-transposition typo

Kuli Jave Copy Prase Clear Preip Isixilosa	
bantwini, iinwele zikhula entioko ubukhulu becala, <mark>Werre</mark> isixa seenwele zomzimba sahlukile kuhlanga nohlanga.	Ignore once Ignore all
	Add to dictionary Suggestions
	kwaye
	Change
	Change all

32 / 43

# Outline

## Motivation

#### 2 Data-driven spellcheckers

- Error detection for isiZulu
- Error detection for isiXhosa
- Error correction for isiZulu and isiXhosa
- Oemo

#### 3 Discussion

#### 4 Conclusions

イロト イポト イヨト イヨト

3

34 / 43

## Discussion

- Cleaned data or noisy data?
- Trigrams on text proper or on non-punctuation-marks?
- Cleaned trigrams or not?

35 / 43

## Discussion

- 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

イロト 不得 とくき とくき とうき

36 / 43

## Discussion

- 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)

37 / 43

## Discussion

- 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

- 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

- 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

## Motivation

#### 2 Data-driven spellcheckers

- Error detection for isiZulu
- Error detection for isiXhosa
- Error correction for isiZulu and isiXhosa
- Oemo

#### 3 Discussion



## Conclusions

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

42 / 43

#### 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.

< ロ > < 同 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < 回 > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ > < □ >

43 / 43

# Thank you!

# Questions?