NLP for African (Nguni) languages

C. Maria Keet

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

Guest lecture NLP course Poznan University of Technology, 10 April 2018
Outline

1 Motivation
   - Context
   - Language ‘crash course’

2 Corpus-based spellcheckers
   - Error detection for isiZulu and isiXhosa
   - Error correction for isiZulu and isiXhosa
   - Discussion

3 Rule-based NLG
   - What is CNL, NLG?
   - Generating basic sentences
   - Language learning exercises

4 Summary
Outline

1 Motivation
   - Context
   - Language ‘crash course’

2 Corpus-based spellcheckers
   - Error detection for isiZulu and isiXhosa
   - Error correction for isiZulu and isiXhosa
   - Discussion

3 Rule-based NLG
   - What is CNL, NLG?
   - Generating basic sentences
   - Language learning exercises

4 Summary
Motivation

- IsiZulu and isiXhosa most widely spoken languages in South Africa by first language speakers
- 23% or about 11 million people (isiZulu), 8 million (isiXhosa)
- Have very limited ICT support
Motivation

- IsiZulu and isiXhosa most widely spoken languages in South Africa by first language speakers
- 23% or about 11 million people (isiZulu), 8 million (isiXhosa)
- Have very limited ICT support
- They use computers for work, social media... (e.g.: avg. 1.5 mobile connection/pp in SA)
- So, NLP for these languages: searching online, spellcheckers, machine translation, speech, etc.
Bantu languages: group of languages spoken in Sub-Saharan Africa

*Bantu* means ‘human’; bit of a laden term, but still used in linguistics

Number of languages varies by who counts (> 200 at least)

Organised in so-called Guthrie zones
Guthrie Zones
Core characteristics

- System of noun classes (besides the I, you (sg.), he/she, we, you (pl.), they)
  - For 3rd pers. sg./pl. nouns (i.e., not pers. pron.): they are classified into a noun class
  - Meinhof identified 23 noun classes (not all of them used)
  - There’s some semantics to them: e.g., NC1 for humans, NC9 for animals, NC15 infinitive nouns
Core characteristics

- System of noun classes (besides the I, you (sg.), he/she, we, you (pl.), they)
  - For 3rd pers. sg./pl. nouns (i.e., not pers. pron.): they are classified into a noun class
  - Meinhof identified 23 noun classes (not all of them used)
  - There’s some semantics to them: e.g., NC1 for humans, NC9 for animals, NC15 infinitive nouns
- Most, but not all, of the languages are agglutinating
  - i.e., what are separate words in, say, English are ‘components’ of a word
  - Ex: titukakimureeterahoganu (Runyankore, Uganda)
  - ‘We have never ever brought it to him’
Core characteristics

- System of noun classes (besides the I, you (sg.), he/she, we, you (pl.), they)
  - For 3rd pers. sg./pl. nouns (i.e., not pers. pron.): they are classified into a noun class
  - Meinhof identified 23 noun classes (not all of them used)
  - There’s some semantics to them: e.g., NC1 for humans, NC9 for animals, NC15 infinitive nouns

- Most, but not all, of the languages are *agglutinating*
  - i.e., what are separate words in, say, English are ‘components’ of a word
  - Ex: `titukakimureeterahoganu` (Runyankore, Uganda)
    - ‘We have never ever brought it to him’
    - `ti tu ka ki mu reet er a ho ga nu`
    - `neg-(NC2 SC)-RM-(NC7 SC)-(NC1 SC)-VR-App-FV-Loc-Emp-Dec`
Core characteristics

- System of noun classes (besides the I, you (sg.), he/she, we, you (pl.), they)
  - For 3rd pers. sg./pl. nouns (i.e., not pers. pron.): they are classified into a noun class
  - Meinhof identified 23 noun classes (not all of them used)
  - There’s some semantics to them: e.g., NC1 for humans, NC9 for animals, NC15 infinitive nouns
- Most, but not all, of the languages are agglutinating
  - i.e., what are separate words in, say, English are ‘components’ of a word
  - Ex: titukakimureeeterahoganu (Runyankore, Uganda)
    ‘We have never ever brought it to him’
    ti tu ka ki mu reet er a ho ga nu
    neg-(NC2 SC)-RM-(NC7 SC)-(NC1 SC)-VR-App-FV-Loc-Emp-Dec
- System of concordial agreement (more about that soon)
<table>
<thead>
<tr>
<th>NC</th>
<th>AU</th>
<th>PRE</th>
<th>Stem (example)</th>
<th>Meaning</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-a</td>
<td>m(u)-</td>
<td>-fana</td>
<td>humans and other animates</td>
<td>umfana</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>ba-</td>
<td>-fana</td>
<td></td>
<td>abafana</td>
</tr>
<tr>
<td>1a</td>
<td>u-o</td>
<td>-</td>
<td>-baba</td>
<td>kinship terms and proper names</td>
<td>ubaba</td>
</tr>
<tr>
<td>2a</td>
<td></td>
<td>-</td>
<td>-baba</td>
<td></td>
<td>obaba</td>
</tr>
<tr>
<td>3a</td>
<td>u-o</td>
<td>-</td>
<td>-shizi</td>
<td>nonhuman</td>
<td>ushizi</td>
</tr>
<tr>
<td>(2a)</td>
<td></td>
<td>-</td>
<td>-shizi</td>
<td></td>
<td>oshizi</td>
</tr>
<tr>
<td>3</td>
<td>u-i</td>
<td>m(u)-</td>
<td>-fula</td>
<td>trees, plants, non-paired body parts</td>
<td>umfula</td>
</tr>
<tr>
<td>4</td>
<td>i-mi</td>
<td>-fula</td>
<td></td>
<td></td>
<td>imifula</td>
</tr>
<tr>
<td>5</td>
<td>i-(li)-</td>
<td>-gama</td>
<td></td>
<td>fruits, paired body parts, and natural phenomena</td>
<td>igama</td>
</tr>
<tr>
<td>6</td>
<td>a-ma</td>
<td>-gama</td>
<td></td>
<td></td>
<td>amagama</td>
</tr>
<tr>
<td>7</td>
<td>i-si</td>
<td>-hlalo</td>
<td></td>
<td>inanimates and manner/style</td>
<td>isihlalo</td>
</tr>
<tr>
<td>8</td>
<td>i-zi</td>
<td>-hlalo</td>
<td></td>
<td></td>
<td>izihlalo</td>
</tr>
<tr>
<td>9a</td>
<td>i-a</td>
<td>-</td>
<td>-rabha</td>
<td>nonhuman</td>
<td>irabha</td>
</tr>
<tr>
<td>(6)</td>
<td></td>
<td>ma-</td>
<td>-rabha</td>
<td></td>
<td>amarabha</td>
</tr>
<tr>
<td>9</td>
<td>i(n)-</td>
<td>-ja</td>
<td></td>
<td>animals</td>
<td>inja</td>
</tr>
<tr>
<td>10</td>
<td>i-zi(n)-</td>
<td>-ja</td>
<td></td>
<td></td>
<td>izinja</td>
</tr>
<tr>
<td>11</td>
<td>u-i</td>
<td>(lu)-</td>
<td>-thi</td>
<td>inanimates and long thin objects</td>
<td>uthi</td>
</tr>
<tr>
<td>(10)</td>
<td></td>
<td>zi(n)-</td>
<td>-thi</td>
<td></td>
<td>izinthi</td>
</tr>
<tr>
<td>14</td>
<td>u-bu-</td>
<td>-hle</td>
<td></td>
<td>abstract nouns</td>
<td>ubuhle</td>
</tr>
<tr>
<td>15</td>
<td>u-ku-</td>
<td>-cula</td>
<td></td>
<td>infinitives</td>
<td>ukucula</td>
</tr>
<tr>
<td>17</td>
<td>ku-</td>
<td></td>
<td></td>
<td>locatives, remote/general</td>
<td></td>
</tr>
</tbody>
</table>
Concordial agreement—example (isiZulu, South Africa)

Abafana abancane bazozithenga izincwadi ezinkulu

'Abafana abancane bazozithenga izincwadi ezinkulu'

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

Abafana abancane bazozithenga izincwadi ezinkulu

'Abafana abancane bazozithenga izincwadi ezinkulu'

The little boys will buy the big books’
Other illustrative examples (isiZulu)

- ‘and’, enumerative: *na-* , phonologically conditioned
  Ex: milk and butter: *ubisi nebhotela* 
  Ex: butter and milk: *ibhotela nobisi*

"is not a": combine NEG SC with PRON, both depend on noun class
Ex: an animal is not a plant: *isilwane asiwona umuthi*
Ex: a plant is not an animal: *umuthi awusona isilwane*

‘all’/’each’ (\(\forall\)), ‘at least one’ (\(\exists\)): quantifiers depend on noun class
Ex: all animals: *zonke izilwane* / all plants: *yonke imithi*
Ex: at least one animal: *isilwane esisodwa* / at least one plant: *umuthi owodwa*

Copulative (to be): depends on first letter of noun: *ng-* for a-, o-, u-,
else: *y-*
Ex: is a dog: *y inja*
Ex: is a grandmother: *ng uqogo*
Other illustrative examples (isiZulu)

- ‘and’, enumerative: *na-*, phonologically conditioned
  - Ex: milk and butter: *ubisi nebhotela* (-a+i=-e-)
  - Ex: butter and milk: *ibhotela nobisi* (-a+u=-o-)

- ‘is not a’: combine NEG SC with PRON, both depend on noun class
  - Ex: an animal is not a plant: *isilwane asiwona umuthi*
  - Ex: a plant is not an animal: *umuthi awusona isilwane*
Other illustrative examples (isiZulu)

- ‘and’, enumerative: *na-*; phonologically conditioned
  - Ex: milk and butter: *ubisi ne*bhotela
  - Ex: butter and milk: *ibhotela no*bisi

- ‘is not a’: combine NEG SC with PRON, both depend on noun class
  - Ex: an animal is not a plant: *isilwane asiwona umuthi*
  - Ex: a plant is not an animal: *umuthi awusona isilwane*

- ‘all’/‘each’ (∀), ‘at least one’ (∃): quantifiers depend on noun class
  - Ex: all animals *zonke izilwane* / all plants *yonke imithi*
  - Ex: at least one animal: *isilwane esisodwa* / at least one plant *umuthi owodwa*

- copulative (to be): depends on first letter of noun: *ng-* for a-, o-, u-, else *y-*
  - Ex: is a dog: *yinja*
  - Ex: is a grandmother: *ngugogo*
Outline

1 Motivation
   • Context
   • Language ‘crash course’

2 Corpus-based spellcheckers
   • Error detection for isiZulu and isiXhosa
   • Error correction for isiZulu and isiXhosa
   • Discussion

3 Rule-based NLG
   • What is CNL, NLG?
   • Generating basic sentences
   • Language learning exercises

4 Summary
Very limited available spellcheckers for use:
- Outdated/software not working anymore (OpenOffice plugin)
- Online one with too many clicks, popups, and ads

1 https://www.spellchecker.net/africa_zulu_spell_checker.html
General issues and aims

- Very limited available spellcheckers for use:
  - Outdated/software not working anymore (OpenOffice plugin)
  - Online one with too many clicks, popups, and ads
- Investigate development of spellchecker for isiZulu
- Find an approach that can be used across (agglutinating) Bantu languages

---

1 https://www.spellchecker.net/africa_zulu_spell_checker.html
What will work best?

- Dictionary approach won’t work due to (theoretically) agglutination and (practically) limited online dictionaries
What will work best?

- Dictionary approach won’t work due to (theoretically) agglutination and (practically) limited online 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 online 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
First iteration [Ndaba et al.(2016)]

1. Use corpus for data
2. Generate n-gram statistics; tried trigrams and quadrigrams; e.g.:
   - original word *ngimbona*
   - trigrams: ngi, gim, imb, ...
   - quadrigrams: ngim, gimb, ...
3. Compute frequencies of the tri/quadrigrams
4. Determine threshold of when a tri/quadrigram would probably be wrong
5. Then when a word is given:
   a. Generate its trigrams
   b. Check probability each trigram
   c. If tri/quadrigram below threshold: flag word as incorrectly spelled
Example language model creation

<table>
<thead>
<tr>
<th>corpus</th>
<th>generate trigrams</th>
<th>unique trigrams</th>
<th>frequency</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>sivela</td>
<td>siv ive vel ela</td>
<td>ala</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td>ngihamba</td>
<td>ngi gih iha ham amb mba</td>
<td>amb</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td>uvelaphi</td>
<td>uve vel ela lap aph phi</td>
<td>aph</td>
<td>2</td>
<td>0.07407407</td>
</tr>
<tr>
<td>ngivala</td>
<td>ngi giv iva val ala</td>
<td>ela</td>
<td>3</td>
<td>0.11111111</td>
</tr>
<tr>
<td>uvelaphi</td>
<td>uve vel ela lap aph phi</td>
<td>gih</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>giv</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ham</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>iha</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>iva</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ive</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>lap</td>
<td>2</td>
<td>0.07407407</td>
</tr>
<tr>
<td></td>
<td></td>
<td>mba</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ngi</td>
<td>2</td>
<td>0.07407407</td>
</tr>
<tr>
<td></td>
<td></td>
<td>phi</td>
<td>2</td>
<td>0.07407407</td>
</tr>
<tr>
<td></td>
<td></td>
<td>siv</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>uve</td>
<td>2</td>
<td>0.07407407</td>
</tr>
<tr>
<td></td>
<td></td>
<td>val</td>
<td>1</td>
<td>0.03703704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>vel</td>
<td>3</td>
<td>0.11111111</td>
</tr>
</tbody>
</table>
Example how this then works in the spellchecker

All trigrams are above the threshold, so it is assumed to have been spelled correctly.

At least one trigram is below the threshold (there is no gni trigram in the language model), so it is probably misspelled, and flagged as such.

A word it was not trained on, but all trigrams are above the threshold, so it is assumed to have been spelled correctly.

At least one trigram is below the threshold (no vea), so it is probably misspelled.

It suggests a correction using more probable trigrams.

Double click on errors to make correction one at a time
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)
Major outcomes

- 89% accuracy (on par with older data [Bosch and Eisele(2005), Prinsloo and de Schryver(2004)])
- The spellchecker performed slightly better with trigrams than with quadrigrams
- Accurate in detecting words that do not occur in the training corpus
- The most updated corpora are preferable
Major outcomes

- 89% accuracy (on par with older data [Bosch and Eisele(2005), Prinsloo and de Schryver(2004)])
- The spellchecker performed slightly better with trigrams than with quadrigrams
- Accurate in detecting words that do not occur in the training corpus
- The most updated corpora are preferable

- 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
Corpus-based spellcheckers

Error correction for isiZulu and isiXhosa

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
  - e.g., with typo: nii
  - intended: 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
  - e.g., with typo: nii
  - intended: 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)
Deletion typo and a suggested correction (running example)

All trigrams are above the threshold, so it is assumed to have been spelled correctly.
At least one trigram is below the threshold (there is no gni trigram in the language model), so it is probably misspelled, and flagged as such.
A word it was not trained on, but all trigrams are above the threshold, so it is assumed to have been spelled correctly.
At least one trigram is below the threshold (no vea), so it is probably misspelled.
It suggests a correction using more probable trigrams.

Double click on errors to make correction one at a time
Error correction–transposition typo (isiZulu)

yebo, ngivela eThekwin
Error correction–transposition typo (isiXhosa)

Ebantwini, iinwele zikhula entloko ubukhulu becala, *wkaye* isixa seenwele zomzimba sahlukile kuhlanga nohlanga.

Double click on errors to make correction one at a time
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
KUBATSHAZWA isiXhulu owesifazane oneminyaka engu-22 edlwengulwa amadoda ayisihlanu eshintshana ngaye eNseleni eMpenge. Umphakathi wakule ndawo usababaza isigameko okuthiwa senzeka ngolwesibili ebusuku.


Umthombo ongathandanga ukudalulwa uthe lolo daba lukhulunyelwaphansi endlaweni njengoba luthathwa njengenhlazo. “Simangazwe yishiluku sabasolwa ngokushulela ngemise beiyiwengula. Kuyacca ukuthi izingane zethu kaziaphile, kusho umthombo.”

KUBATSHAZWA isihluku owesifazane oneminyaka engu–22 edlwengulwa amadoda ayisihlanu eshintshana ngaye eNseleni eMpange. Umpakathi wakule ndawo usababaza isigameko okuthiwa senzeke ngolwesibili ebusuku.

Kuthiwa owesifazane ubanjwe amadoda athe ayamphereleza esegodu ka esuka eJoyintini. Ngokuthola kweSoledzwe lo wesifazane kade "czijabulisa" eJoyintini okuthiwa yiFikile khona eNseleni, Kuthiwa ubecele umfowabo ukuba amhlangoabeze. Amadoda amahlanu okuthiwa amenzakalisile kxhela ukuthi aye kuye amthela ukuthi ayamazi umfowabo ngakho azompheleleza ngoba ubengakafiki.

Lo wesifazane kuthiwa uhambabe nabasolwa okuthe besendleleni bamnikela. Kuthiwa bahambe naye ngenkani bayomfaka endlini abafike bamdlwengula kuyona. Eksenzi kuthiwa uhunyekile wayobika ngokumhlele.

Okhulumela amaphoyisa KwaZulu-Natal uColonel Jay Naicker ukuqinisekile ukuthi amaphoyisa aseMpange aphenya icala lokudlwengulwa kowesifazane.


Double click on errors to make correction one at a time
Wagqibela nini ukuza kwenza uvavanyo lomzima?
Ndagqibela ukwenza uvavanyo lwam kwiminyaka emibini edlulileyo.
Ubukhe wana olo olunye uvavanyo kutsha nje?
Ukusebenza kwegazi, iEKG okanye iultra-sound
Ndingatsho, bendikhe ndenza i-X-ray ezimbalwa kwaggirha wamazinyo.
Ubuhieli uzuva njani ngokuthe jikelele?
Akukho zikhala zo, nyabani.
Uwunakukho unyuse umkhono wasekholo?
Ndinqwenela ukuthatha ifutha lakho legazi.
Ngama-120 angapehe kuwama 80
Awubonakali ngathi uyebe kakhulu, intle loo nto
Ingaba uzilolanga njalo?
Hayi, andingenjayo.
Ukuba ndinyuka ndibaleka izitepsi, kundithatha ixesha ukuba ndifumane ukuphafumla kwam kwakhona.
Kudingeka ndiphume ngokongezeleleleayo.
Ing Alicebo elihle elo.
Injanjana yona indlela otya ngayo?
Ndicinga ndiya ngendlela efanelekeleyo.
Uyazi, ndiba nehamburger elo xesha nelo xesha, kodwa ngokuthe jikelele, ndiya ukutya okufanelekeleyo.
Ngoku, ndiza kumamela intliziyo yakho.
Yhu, iyabanda!
Musa ukukhathazeka, vistingoscope yam je kuphela.
Ngoku, phumyilela ngaphakathi uze uwubambela umphefumlo wakho.
Ndicela unyuse ishethi yakho, uphafumle kakhulu.
Yonke into ivakala kakhule.
Masikhe sijonge umqala wakho.
Ndicela uvule kakhulu uze uthi 'ah'.
Corpus-based spellcheckers

Error correction for isiZulu and isiXhosa

isiXhosa text, isiZulu model

Wagqibela nini ukuza uvanyo lomzimba?
Ndaggibela ukwenza uvanyo lwam kwiminyaka emitini edlulileyo.
Ubukhe wănalo olunye uvanyo kutsha nje?
Ukusebenza kwegazi, IFKC okanye illra-sound
Ndingatsho, bendikhe ndenza i-X-rays ezimbalwa kwaggirha wamazinyo.
Ubuhleli uziva njani ngokuthe jikelele?
Akukho zikhalaazo, nyhani.
Uwunakukhe unyuso umkhono wasekhoholo?
Ndinqweni ukuthatha ifuthu lakho legazi.
ngama-120 angaphezu kwama-80
Awubonakali ngathi unyebhe kaakhulu, intle loo nto
Ingaba uzilolonga njalo?
Hayi, andingetsho

Ukuba ndinyuka ndibaleka izitupisi, kundithatha ixesha ukuba ndifumane ukuphefumla kwam kwakhona.
Kudingeka ndiphume ngokongezelelewyo.
Ingalicebo eihle elo.
Injani yona indlela otya ngayo?
Ndicinga ndiya ngendlela efanelekileyo.
Uyazi, ndiba nehamburger elo xesha neloxesha, kodwa ngokuthe jikelele, ndiya ukutya okufanelekileyo.
Ngoku, ndiza kumamelu intliziyo yakho.
Yhu, iyabanda!
Musa ukukhathazeka, vistethoscope yam je kuphela.
Ngoku, pherentlela ngaphakathi uze uubambhe umphefumlo wakho.
Ndicela unyuse isheti yakho, uphefumele kakhu.
Yonke into ivakala kakhile.
Masikhe sijonge umqala wakho.
Ndicela uvela kakhu uze uthi 'ah'.

Double click on errors to make correction one at a time
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

1 Motivation
   ● Context
   ● Language ‘crash course’

2 Corpus-based spellcheckers
   ● Error detection for isiZulu and isiXhosa
   ● Error correction for isiZulu and isiXhosa
   ● Discussion

3 Rule-based NLG
   ● What is CNL, NLG?
   ● Generating basic sentences
   ● Language learning exercises

4 Summary
Short answer

- **C**ontrolled **N**atural **L**anguage: constrain the grammar/vocabulary of a natural language
- **N**atural **L**anguage **G**eneration: generate natural language text from structured data, information, or knowledge
Natural language interfaces with some CNL or NLG

- Many tools, webpages, etc. with some natural language component
- Querying of information in natural language (cf. a query language SQL, SPARQL)
- Business rules typically specified in a natural language
- etc.
Example: iCal calendar entry with canned text
Example: Saadiq Moolla’s mobile healthcare app

Chest Pain

Have you had any recent pain in your chest? - Uke waba nobuhlunlu esifubeni maduzane?

Does the pain radiate to your jaw, neck or arm? - Engabe ubuhlunlu bakho bujikeleza emihlathini, emqaleni noma nasezingalweni?

Does anything precipitate or relieve the pain? - Ingabe ikhona into eyenza ubuhlunlu buqhubeke noma eyehlisa ubuhlunlu?

Dyspnoea

Have you had any recent pain in your chest? - Ingaba kutshanje ukhe weva iintlungu esifubeni?

Does the pain radiate to your jaw, neck or arm? - Ingaba iintlungu zinwenwela emihlathini, entanyeni okanye engalweni?

Does anything precipitate or relieve the pain? - Ingabe ikhona into eyenza ubuhlunlu buqhubeke noma eyehlisa ubuhlunlu?
Example: Query formulation with Quelo
[Franconi et al. (2010)]

Pictures from: Quelo © The IESD Challenge 2012
Demo at: http://krdbapp.inf.unibz.it:8080/quelo/
Example: Business rules and conceptual data models

Each Course is taught by at least one Professor
Each Professor teaches at least one Course
The ‘NLG pipeline’

1. What structured data/info/knowledge do you want to put into NL sentences?
2. In what order should it be presented?
3. Which messages to put together into a sentence?
4. Which words and phrases will it use for each domain concept and relation?
5. Which words or phrases to select to identify domain entities?
6. Use grammar rules to produce syntactically, morphologically, and orthographically correct (and is also meaningful)
NLG, principal approaches to generate the text

- Canned text
- Templates
  - but also other languages [Jarrar et al.(2006)]
- Grammar engines, such as [Kuhn(2013)], Grammatical Framework (http://www.grammaticalframework.org/), SimpleNLG
NLG, principal approaches to generate the text

- Canned text
- Templates
  - but also other languages [Jarrar et al.(2006)]
- Grammar engines, such as [Kuhn(2013)], Grammatical Framework (http://www.grammaticalframework.org/), SimpleNLG

⇒ CNL, NLG
Business rules/conceptual data models and logic reconstruction

BR: **Each** Course is taught by **at least one** Professor

FOL: $\forall x \ (\text{Course}(x) \rightarrow \exists y \ (\text{is\_taught\_by}(x, y) \land \text{Professor}(y)))$

DL: $\text{Course} \sqsubseteq \exists \text{is\_taught\_by. Professor}$
Example of templates

```xml
<Constraint xsi:type="Mandatory">
  <Text>-[Mandatory] Cada</Text>
  <Object index="0"/>
  <Text>Debe</Text>
  <Role index="0"/>
  <Text>al menos un(a)</Text>
  <Object index="1"/>
</Constraint>

<Constraint xsi:type="Mandatory">
  <Text>-[Mandatory] Each</Text>
  <Object index="0"/>
  <Text>Must</Text>
  <Role index="0"/>
  <Text>at least one</Text>
  <Object index="1"/>
</Constraint>
```

for a large fragment of ORM, and 11 languages [Jarrar et al.(2006)]
Example of templates

for a large fragment of ORM, and 11 languages [Jarrar et al.(2006)]
Example of templates

for a large fragment of ORM, and 11 languages [Jarrar et al.(2006)]
Example of templates

for a large fragment of ORM, and 11 languages [Jarrar et al.(2006)]
NL Grammars, illustration

\[
\begin{align*}
\text{Sentence} & \rightarrow \text{NounPhrase} \mid \text{VerbPhrase} \\
\text{NounPhrase} & \rightarrow \text{Adjective} \mid \text{NounPhrase} \\
\text{NounPhrase} & \rightarrow \text{Noun} \\
\text{Noun} & \rightarrow \text{car} \mid \text{train} \\
\text{Adjective} & \rightarrow \text{big} \mid \text{broken} \\
\end{align*}
\]

(and complexity of the grammar)
Can the template-based approach be used also for isiZulu?
Question

- Can the template-based approach be used also for isiZulu?
  - If so, create those templates
  - If not, start with basics for a grammar engine
Question

- Can the template-based approach be used also for isiZulu?
  - If so, create those templates
  - If not, start with basics for a grammar engine

- Use a practically useful language to benefit both ICT and linguists and, possibly, some subject domain (e.g., medicine)

A logic foundation for isiZulu knowledge-to-text

- Roughly OWL 2 EL
- OWL 2 EL is a W3C-standardised profile of OWL 2
- Tools, ontologies in OWL 2 (notably SNOMED CT)
\textbf{ALC syntax}

- \textit{Concepts} denoting entity types/classes/unary predicates/universals, including top $\top$ and bottom $\bot$;
- \textit{Roles} denoting relationships/associations/n-ary predicates/properties;
- \textit{ Constructors}: and $\sqcap$, or $\sqcup$, and not $\neg$; quantifications for all $\forall$ and exists $\exists$
- \textit{Complex concepts} using constructors: Let $C$ and $D$ be concept names, $R$ a role name, then
  - $\neg C$, $C \sqcap D$, and $C \sqcup D$ are concepts, and
  - $\forall R.C$ and $\exists R.C$ are concepts
- \textit{Individuals}
**ALC** semantics

- **domain of interpretation**, and an **interpretation**, where:
  - Domain $\Delta$ is a non-empty set of objects
  - Interpretation: $\cdot^I$ is the *interpretation function*, domain $\Delta^I$
    - $\cdot^I$ maps every concept name $A$ to a subset $A^I \subseteq \Delta^I$
    - $\cdot^I$ maps every role name $R$ to a subset $R^I \subseteq \Delta^I \times \Delta^I$
    - $\cdot^I$ maps every individual name $a$ to elements of $\Delta^I$: $a^I \in \Delta^I$
  - Note: $\top^I = \Delta^I$ and $\bot^I = \emptyset$
- $(\lnot C)^I = \Delta^I \setminus C^I$
- $(C \cap D)^I = C^I \cap D^I$
- $(C \cup D)^I = C^I \cup D^I$
- $(\forall R.C)^I = \{ x \mid \forall y. R^I(x, y) \rightarrow C^I(y) \}$
- $(\exists R.C)^I = \{ x \mid \exists y. R^I(x, y) \land C^I(y) \}$
Universal Quantification

- Consider here only the universal quantification at the start of the concept inclusion axiom (‘nominal head’)
- ‘all’/‘each’ uses -onke, prefixed with the oral prefix of the noun class of that first noun (OWL class/DL concept) on lhs of ⊑

(U1) Boy ⊑ ...
   wonke umfana ...
   bonke abafana ...

(U2) Phone ⊑ ...
   lonke ifoni ...
   onke amafoni ...

('each boy...'; u- + -onke)
('all boys...'; ba- + -onke)
('each phone...'; li- + -onke)
('all phones...'; a- + -onke)
<table>
<thead>
<tr>
<th>NC</th>
<th>QC&lt;sub&gt;oral+onke&lt;/sub&gt;</th>
<th>QC&lt;sub&gt;nke&lt;/sub&gt;</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QC&lt;sub&gt;dwa&lt;/sub&gt;</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
</tr>
<tr>
<td>NC</td>
<td>QC (all)</td>
<td>NEG SC</td>
<td>PRON</td>
<td>RC</td>
<td>QC_dwa</td>
<td>EC</td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>--------------------------</td>
<td>--------</td>
<td>------</td>
<td>----</td>
<td>---------</td>
<td>-----</td>
<td></td>
</tr>
<tr>
<td></td>
<td>QC_oral+onke</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>QC&lt;sub&gt;oral&lt;/sub&gt;</td>
<td>QC&lt;sub&gt;all&lt;/sub&gt;</td>
<td>QC&lt;sub&gt;nke&lt;/sub&gt;</td>
<td>NEG SC</td>
<td>PRON</td>
<td>RC</td>
<td>QC&lt;sub&gt;dwa&lt;/sub&gt;</td>
</tr>
<tr>
<td>-----</td>
<td>------------------</td>
<td>----------------</td>
<td>---------------</td>
<td>-------</td>
<td>-------</td>
<td>-----</td>
<td>----------------</td>
</tr>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>ye-</td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>ye-</td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>ye-</td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
</tr>
</tbody>
</table>
Subsumption

- Two different ways of carving up the nouns to determine which rules apply: semantic and syntactic
- Need to choose between
  - singular and plural
  - with or without the universal quantification voiced
  - generic or determinate

(S1) MedicinalHerb ⊑ Plant
  ikhambi ngumuthi
  amakhambi yimithi
  wonke amakhambi ngumuthi
  (‘medicinal herb is a plant’) (‘medicinal herbs are plants’) (‘all medicinal herbs are a plant’)

(S2) Giraffes ⊑ Animals
  izindlulamithi yizilwane
  (‘giraffes are animals’; generic)

(S3) Cellphone ⊑ Phone
  Umakhalekhukhwini uyifoni
  (‘cellphone is a phone’; determ.)
Possible subsumption patterns

a. $N_1 \text{ copulative } ng/y \text{ depending on first letter of } N_2 > N_2$.  

b. $\text{plural of } N_1 > \text{copulative } ng/y \text{ depending on first letter of plural of } N_2 > \text{plural of } N_2$.  

c. $\text{All-concord for } NC_x > \text{onke } \text{plural of } N_1, \text{being of } NC_x > \text{copulative } ng/y \text{ depending on first letter of } N_2 > N_2$.  

Subsumption: adding negation

- Need to choose between
  - singular and plural, and with or without the universal quantification voiced
- Copulative is omitted
- Combines the negative subject concord (NEG SC) of the noun class of the first noun (aku-) with the pronomial (PRON) of the noun class of second noun (-yona)

$\text{(SN1)} \text{ Cup } \not\subseteq \neg \text{ Glass}$

- indebe aku-yona ingilazi
- zonke izindebe aziyona ingilazi

('cup not a glass')

('all cups not a glass')
## Rule-based NLG Generating basic sentences

<table>
<thead>
<tr>
<th>NC</th>
<th>QCoral+onke</th>
<th>QCnke</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QCdwa</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yoa-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
</tr>
</tbody>
</table>
## Rule-based NLG
### Generating basic sentences

<table>
<thead>
<tr>
<th>NC</th>
<th>QC (all)</th>
<th>QC&lt;sub&gt;oral+onke&lt;/sub&gt;</th>
<th>QC&lt;sub&gt;nke&lt;/sub&gt;</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QC&lt;sub&gt;dwa&lt;/sub&gt;</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>ab-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>ab-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>3a (2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>ab-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>el-</td>
<td>lo-</td>
<td>li-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>9a (6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
<td></td>
</tr>
<tr>
<td>11 (10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>ob-</td>
<td>bo-</td>
<td>bu-</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>ok-</td>
<td>zo-</td>
<td>ku-</td>
<td></td>
</tr>
<tr>
<td>NC</td>
<td>QC (all)</td>
<td>NEG SC</td>
<td>PRON</td>
<td>RC</td>
<td>QCdwa</td>
<td>EC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>----------</td>
<td>--------</td>
<td>------</td>
<td>----</td>
<td>-------</td>
<td>----</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QC&lt;sub&gt;oral+onke&lt;/sub&gt;</td>
<td>QC&lt;sub&gt;nke&lt;/sub&gt;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>bo-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>bo-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>bo-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>bo-</td>
<td>wo-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>ma-</td>
<td>ma-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>z-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>z-</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>abu-</td>
<td>bo-</td>
<td>bu-</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>ku-</td>
<td>ku-</td>
<td></td>
</tr>
</tbody>
</table>
Possible negation (disjointness) patterns

a. $<N_1 \text{ of } NC_x> <\text{NEG SC of } NC_x><\text{PRON of } NC_y> <N_2 \text{ of } NC_y>.$

b. $<\text{All-concord for } NC_x> \text{onke } <\text{plural } N_1, \text{ being of } NC_x> <\text{NEG SC of } NC_x><\text{PRON of } NC_y> <N_2 \text{ with } NC_y>.$
Existential Quantification

(E1) Giraffe ⊑ ∃eats.Twig
    yonke indlulamithi idla ihlamvana elilodwa
    zonke izindlulamithi zidla ihlamvana elilodwa

('each giraffe eats at least one twig')
('all giraffes eat at least one twig')

a. <All-concord for NCₓ>onke <pl. N₁, is in NCₓ> <conjugated verb>
    <N₂ of NCᵧ> <RC for NCᵧ> <QC for NCᵧ>dwa.
<table>
<thead>
<tr>
<th>NC</th>
<th>QC (all)</th>
<th>QC&lt;sub&gt;oral+onke&lt;/sub&gt;</th>
<th>QC&lt;sub&gt;nke&lt;/sub&gt;</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QC&lt;sub&gt;dwa&lt;/sub&gt;</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yon-</td>
<td>mi-</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yon-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yon-</td>
<td>yi-</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
<td></td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
<td></td>
</tr>
</tbody>
</table>
### Rule-based NLG: Generating basic sentences

<table>
<thead>
<tr>
<th>NC</th>
<th>QC (all)</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QC Ωwa</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>a-</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>a-</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zì-</td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zì-</td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zu-</td>
<td>ku-</td>
</tr>
</tbody>
</table>
## Rule-based NLG

Generating basic sentences

<table>
<thead>
<tr>
<th>NC</th>
<th>QC\textsubscript{oral+onke}</th>
<th>QC\textsubscript{nke}</th>
<th>NEG SC</th>
<th>PRON</th>
<th>RC</th>
<th>QC\textsubscript{dwa}</th>
<th>EC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>yena</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>2a</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3a</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>aka-</td>
<td>wona</td>
<td>o-</td>
<td>ye-</td>
<td>mu-</td>
</tr>
<tr>
<td>(2a)</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>aba-</td>
<td>bona</td>
<td>aba-</td>
<td>bo-</td>
<td>ba-</td>
</tr>
<tr>
<td>3</td>
<td>u-onke → wonke</td>
<td>wo-</td>
<td>awu-</td>
<td>wona</td>
<td>o-</td>
<td>wo-</td>
<td>mu-</td>
</tr>
<tr>
<td>4</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>li-onke → lonke</td>
<td>lo-</td>
<td>ali-</td>
<td>lona</td>
<td>eli-</td>
<td>lo-</td>
<td>li-</td>
</tr>
<tr>
<td>6</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>si-onke → sonke</td>
<td>so-</td>
<td>asi-</td>
<td>sona</td>
<td>esi-</td>
<td>so-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>(6)</td>
<td>a-onke → onke</td>
<td>o-</td>
<td>awa-</td>
<td>wona</td>
<td>a-</td>
<td>wo-</td>
<td>ma-</td>
</tr>
<tr>
<td>9</td>
<td>i-onke → yonke</td>
<td>yo-</td>
<td>ayi-</td>
<td>yona</td>
<td>e-</td>
<td>yo-</td>
<td>yi-</td>
</tr>
<tr>
<td>10</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>11</td>
<td>lu-onke → lonke</td>
<td>lo-</td>
<td>alu-</td>
<td>lona</td>
<td>olu-</td>
<td>lo-</td>
<td>lu-</td>
</tr>
<tr>
<td>(10)</td>
<td>zi-onke → zonke</td>
<td>zo-</td>
<td>azi-</td>
<td>zona</td>
<td>ezi-</td>
<td>zo-</td>
<td>zi-</td>
</tr>
<tr>
<td>14</td>
<td>ba-onke → bonke</td>
<td>bo-</td>
<td>abu-</td>
<td>bona</td>
<td>obu-</td>
<td>bo-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>ku-onke → konke</td>
<td>zo-</td>
<td>aku-</td>
<td>khona</td>
<td>oku-</td>
<td>zo-</td>
<td>ku-</td>
</tr>
</tbody>
</table>
Example

- $\forall x \ (\text{Professor}(x) \rightarrow \exists y \ (\text{teaches}(x, y) \land \text{Course}(y)))$
- Professor $\sqsubseteq \exists x \ \text{teaches}.\text{Course}$
- Each Professor teaches at least one Course
Example

- $\forall x \ (\text{Solwazi}(x) \rightarrow \exists y \ (\text{ufundisa}(x, y) \land \text{Isifundo}(y)))$
- $\text{Solwazi} \sqsubseteq \exists \ \text{ufundisa}. \text{Isifundo}$
- $?$
\[ \forall x \; (uSolwazi(x) \rightarrow \exists y \; (ufundisa(x, y) \land Isifundo(y))) \]
\[ uSolwazi \subseteq \exists \; ufundisa.\; Isifundo \]
Rule-based NLG
Generating basic sentences

∀x (uSolwazi(x) → NC, AU, PRE, uSolwazi ∈ ʃ ufunc)

<table>
<thead>
<tr>
<th>NC</th>
<th>AU</th>
<th>PRE</th>
<th>QCoral+onke</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-</td>
<td>m(u)-</td>
<td>u-onke → wonke</td>
</tr>
<tr>
<td>2</td>
<td>a-</td>
<td>ba-</td>
<td>ba-onke → bonke</td>
</tr>
<tr>
<td>1a</td>
<td>u-</td>
<td></td>
<td>u-onke → wonke</td>
</tr>
<tr>
<td>2a</td>
<td>o-</td>
<td></td>
<td>ba-onke → bonke</td>
</tr>
<tr>
<td>3</td>
<td>u-</td>
<td></td>
<td>u-onke → wonke</td>
</tr>
<tr>
<td>4</td>
<td>m(u)-</td>
<td></td>
<td>ba-onke → bonke</td>
</tr>
<tr>
<td>5</td>
<td>i-</td>
<td>(li)-</td>
<td>u-onke → wonke</td>
</tr>
<tr>
<td>6</td>
<td>a-</td>
<td>ma-</td>
<td>i-onke → yonke</td>
</tr>
<tr>
<td>7</td>
<td>i-</td>
<td>si-</td>
<td>li-onke → lonke</td>
</tr>
<tr>
<td>8</td>
<td>i-</td>
<td>zi-</td>
<td>a-onke → onke</td>
</tr>
<tr>
<td>9</td>
<td>i(n)-</td>
<td></td>
<td>si-onke → sonke</td>
</tr>
<tr>
<td>10</td>
<td>i-</td>
<td></td>
<td>zi-onke → zonke</td>
</tr>
<tr>
<td>11</td>
<td>u-</td>
<td>(lu)-</td>
<td>i-onke → yonke</td>
</tr>
<tr>
<td>14</td>
<td>u-</td>
<td>bu-</td>
<td>a-onke → onke</td>
</tr>
<tr>
<td>15</td>
<td>u-</td>
<td>ku-</td>
<td>i-onke → yonke</td>
</tr>
<tr>
<td>17</td>
<td>ku-</td>
<td></td>
<td>zi-onke → zonke</td>
</tr>
</tbody>
</table>

Bonke oSolwazi

Look-up NC
Pluralise
For-all

Bonke oSolwazi
\[ \forall x (uSolwazi(x) \rightarrow \exists y (ufundisa(x, y) \land Isifundo(y))) \]

\[ uSolwazi \subseteq \exists (ufundisa) \]

... for relevant NC. Here:

- ngi-
- u-
- u-
- si-
- ni-
- ba-

AlgoConjugate

Bonke oSolwazi bafundisa
∀x (uSolwazi(x) → ∃y (ufundisa(x, y) ∧ Isifundo(y)))

Bonke oSolwazi bafundisa Isifundo
**Rule-based NLG**

### Generating basic sentences

<table>
<thead>
<tr>
<th>NC</th>
<th>AU</th>
<th>PRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>u-</td>
<td>m(u)-</td>
</tr>
<tr>
<td>2</td>
<td>a-</td>
<td>ba-</td>
</tr>
<tr>
<td>1a</td>
<td>u-</td>
<td>-</td>
</tr>
<tr>
<td>2a</td>
<td>o-</td>
<td>-</td>
</tr>
<tr>
<td>3a</td>
<td>u-</td>
<td>m(u)-</td>
</tr>
<tr>
<td>4</td>
<td>i-</td>
<td>mi-</td>
</tr>
<tr>
<td>5</td>
<td>i-</td>
<td>(li)-</td>
</tr>
<tr>
<td>6</td>
<td>a-</td>
<td>ma-</td>
</tr>
<tr>
<td>7</td>
<td>i-</td>
<td>si-</td>
</tr>
<tr>
<td>8</td>
<td>i-</td>
<td>zi-</td>
</tr>
<tr>
<td>9a</td>
<td>i-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>i(n)</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>i-</td>
<td>zi(n)</td>
</tr>
<tr>
<td>11</td>
<td>u-</td>
<td>(lu)-</td>
</tr>
<tr>
<td>12</td>
<td>i-</td>
<td>zi(n)</td>
</tr>
<tr>
<td>14</td>
<td>u-</td>
<td>bu-</td>
</tr>
<tr>
<td>15</td>
<td>u-</td>
<td>ku-</td>
</tr>
<tr>
<td>17</td>
<td>ku-</td>
<td>-</td>
</tr>
</tbody>
</table>

**look-up NC**

**get RC**

**get QC**

**add -dwa**

---

**Bonke oSolwazi bafundisa Isifundo esisodwaba**
How to evaluate?

- Typical way of evaluating: ask linguists and/or intended target group
- Questions depend on what you want to know; e.g.,
  - Does the text capture the semantics adequately?
  - Must it really be grammatically correct or is understandable also acceptable?
  - Compared against alternate representation (figures, tables) or human-authored text?
How to evaluate?

- Typical way of evaluating: ask linguists and/or intended target group
- Questions depend on what you want to know; e.g.,
  - Does the text capture the semantics adequately?
  - Must it really be grammatically correct or is understandable also acceptable?
  - Compared against alternate representation (figures, tables) or human-authored text?
- Survey, asked linguists and non-linguists for their preferences
- 10 questions pitting the patterns against each other
- Online, with isiZulu-localised version of Limesurvey
The NLG algorithms can be used elsewhere

- Paper-based language learning exercises
- Exercise books have a lot of exercises on ‘give plural noun’, ‘complete verb’ etc
The NLG algorithms can be used elsewhere

- Paper-based language learning exercises
- Exercise books have a lot of exercises on ‘give plural noun’, ‘complete verb’ etc
- Our algorithms already can do that!
- Reuse the algorithms to pluralise and conjugate
- Proof of concept tool, tried to use both NLP (corpus, POS tagger) and the grammar engine of NLG
Examples of the CNL it uses

- **Pluralise subject**

  **Q:** *Umfowethu bayaphuza*
  **A:** *Abafowethu bayaphuza*

  
  
  
  [prefixSG+stem] [PLSC+VerbRoot+FV]
  [prefixPL+stem] [PLSC+VerbRoot+FV]

  
  
  
  
  Negate the verb

  **Q:** *Batotoba*
  **A:** *Abatotobi*

  
  
  
  
  [PLSC+VerbRoot+FV]
  [PLNEGSC+VerbRoot+NEGFV]

  
  
  
  
  Possible to combine components for new exercises

  
  
  
  
  [prefixSG+stem] [SGSC+VerbRoot+FV] [prefixSG+stem]
  [prefixPL+stem] [PLNEGSC+VerbRoot+NEGFV] [prefixPL+stem]

  
  
  
  
  Q:** umfowethu usula inkomishi**
  A:** abafowethu abasuli izinkomishi**

  
  
  
  
  '*(my) brother washes the cup'*

  
  
  
  
  '*(my) brothers do not wash the cups'*
Examples of the CNL it uses

- **Pluralise subject**
  
  **Q:** *Umfowethu bayaphuza*  
  **A:** *Abafowethu bayaphuza*  
  
  
  [prefixSG+stem] [PLSC+VerbRoot+FV]  
  [prefixPL+stem] [PLSC+VerbRoot+FV]

- **Negate the verb**

  **Q:** *Batotoba*  
  **A:** *Abatotobi*  
  
  [PLSC+VerbRoot+FV]  
  [PLNEGSC+VerbRoot+NEGFV]
Examples of the CNL it uses

- **Pluralise subject**
  
  Q:  *Umfowethu bayaphuza*
  
  A:  *Abafowethu bayaphuza*

  \[[prefixSG+stem] [PLSC+VerbRoot+FV] \]
  \[[prefixPL+stem] [PLSC+VerbRoot+FV] \]

- **Negate the verb**
  
  Q:  *Batotoba*
  
  A:  *Abatotobi*

  \[[PLSC+VerbRoot+FV] \]
  \[[PLNEGSC+VerbRoot+NEGFV] \]

- **Possible to combine components for new exercises**

  \[[prefixSG+stem] [SGSC+VerbRoot+FV] [prefixSG+stem] \]
  \[[prefixPL+stem] [PLNEGSC+VerbRoot+NEGFV] [prefixPL+stem] \]

  Q:  *umfowethu usula inkomishi* ‘(my) brother washes the cup’
  
  A:  *abafowethu abasuli izinkomishi* ‘(my) brothers do not wash the cups’
Outline

1. Motivation
   - Context
   - Language ‘crash course’

2. Corpus-based spellcheckers
   - Error detection for isiZulu and isiXhosa
   - Error correction for isiZulu and isiXhosa
   - Discussion

3. Rule-based NLG
   - What is CNL, NLG?
   - Generating basic sentences
   - Language learning exercises

4. Summary
Some natural language understanding, some generation

- N-grams and learning from a corpus (spellchecker)
- Corpus affects how well the tool performs (spellchecker, isiZulu CALL)
- Templates inapplicable to isiZulu due to its grammar (OWL verbalisation), hence a tailor-made grammar engine
- NLG algorithms generic and modularised in the sense that they can be reused in other tools (CALL exercises)

- Not addressed much now, but no less important: underresourced language
Sonja E. Bosch and Roald Eisele.
The effectiveness of morphological rules for an isiZulu spelling checker.

M. Curland and T. Halpin.
Model driven development with NORMA.
Los Alamitos, Hawaii.

Enrico Franconi, Paolo Guagliardo, and Marco Trevisan.
An intelligent query interface based on ontology navigation.
Hong Kong, February 2010.

Norbert E. Fuchs, Kaarel Kaljurand, and Tobias Kuhn.
Discourse Representation Structures for ACE 6.6.
Technical Report ifi-2010.0010, Department of Informatics, University of Zurich, Zurich, Switzerland, 2010.

Mustafa Jarrar, C. Maria Keet, and Paolo Dongilli.
Multilingual verbalization of ORM conceptual models and axiomatized ontologies.

C. M. Keet and L. Khumalo.
Toward a knowledge-to-text controlled natural language of isiZulu.
C. Maria Keet and Langa Khumalo.

Toward verbalizing logical theories in isiZulu.
20-22 August 2014, Galway, Ireland.

C. Maria Keet and Langa Khumalo.

Basics for a grammar engine to verbalize logical theories in isiZulu.
August 18-20, 2014, Prague, Czech Republic.

Langa Khumalo.

Advances in developing corpora in African languages.

Tobias Kuhn.

A principled approach to grammars for controlled natural languages and predictive editors.

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

The effects of a corpus on isizulu spellcheckers based on n-grams.
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.  

A comparison of three controlled natural languages for OWL 1.1.  
Washington, DC, USA metropolitan area, on 1-2 April 2008.

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach.  
Ukwabelana – an open-source morphological zulu corpus.  
Beijing.

Allan Third, Sandra Williams, and Richard Power.  
OWL to English: a tool for generating organised easily-navigated hypertexts from ontologies.  
poster/demo paper, Open University UK, 2011.  
Thank you!

Questions?
Spellchecker with contributions from:

- IsiZulu Linguist: Langa Khumalo
- 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