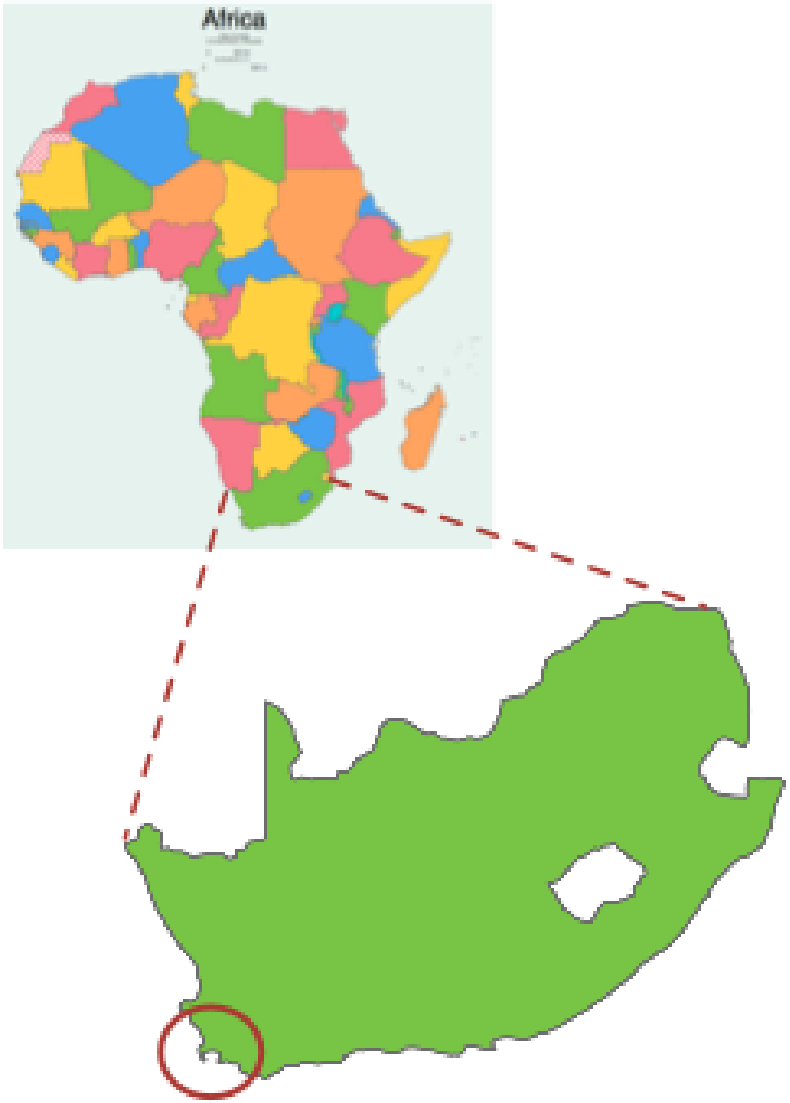
NLP for African (Nguni) languages

C. Maria Keet

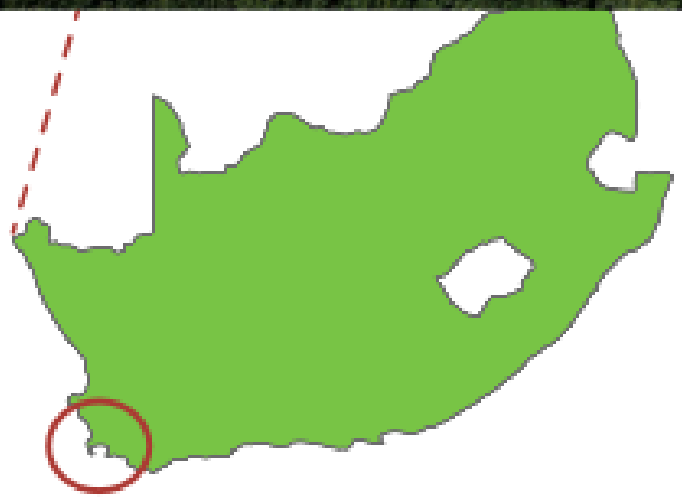
Department of Computer Science
University of Cape Town, South Africa
mkeet@cs.uct.ac.za

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

CS@UCT



CS@UCT



CS@UCT

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

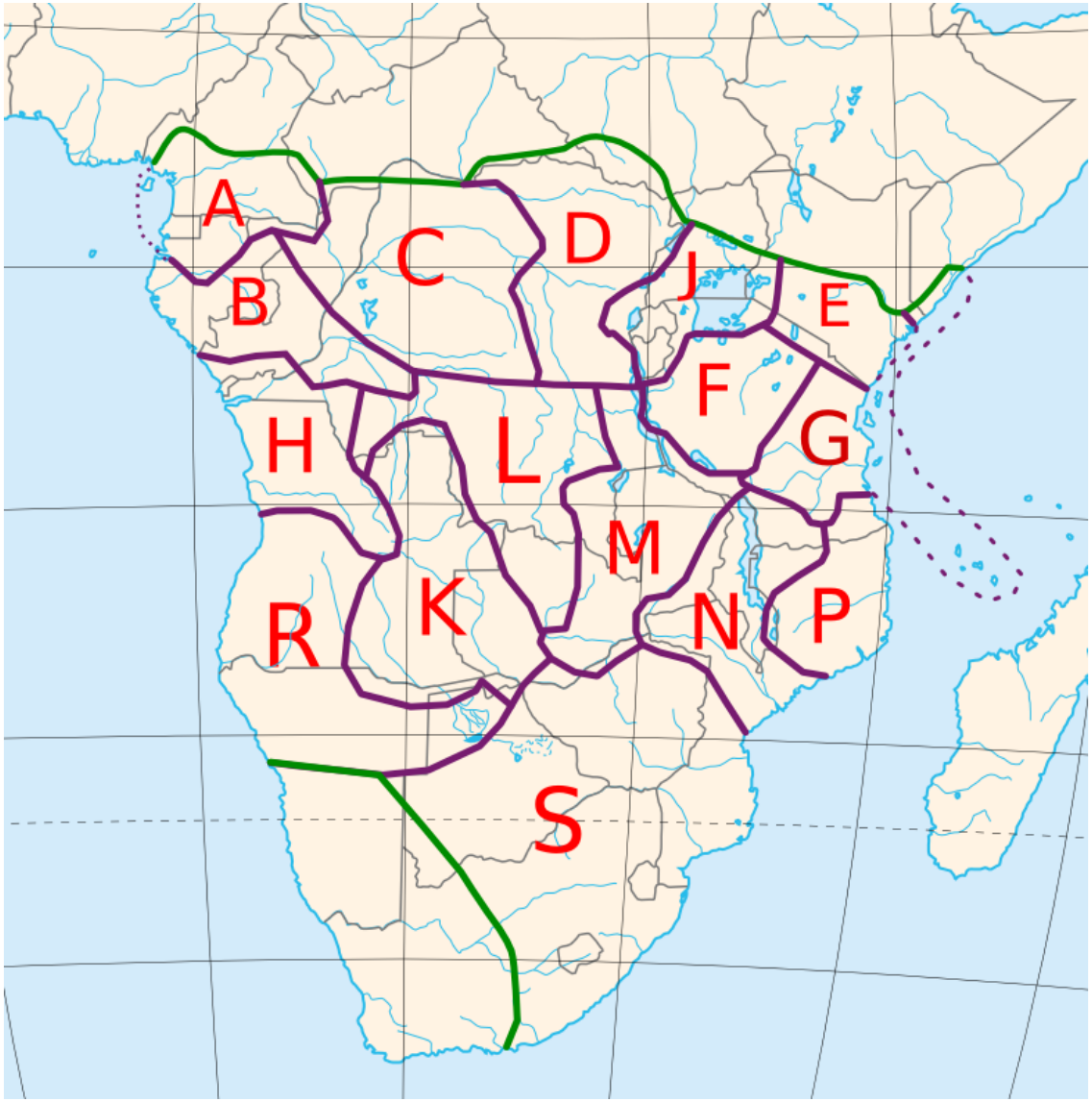
Motivation

- 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.

Basics

- 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: `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
- System of concordial agreement (more about that soon)

NC	AU	PRE	Stem (example)	Meaning	Example	
1 2	u- a-	m(u)- ba-	-fana -fana	humans and other animates	umfana abafana	boy boys
1a 2a	u- o-	- -	-baba -baba	kinship terms and proper names	ubaba obaba	father fathers
3a (2a)	u- o-	- -	-shizi -shizi	nonhuman	ushizi oshizi	cheese cheeses
3 4	u- i-	m(u)- mi-	-fula -fula	trees, plants, non-paired body parts	umfula imifula	river rivers
5 6	i- a-	(li)- ma-	-gama -gama	fruits, paired body parts, and natural phenomena	igama amagama	name names
7 8	i- i-	si- zi-	-hlalo -hlalo	inanimates and manner/ style	isihlalo izihlalo	chair chairs
9a (6)	i- a-	- ma-	-rabha -rabha	nonhuman	irabha amarabha	rubber rubbers
9 10	i(n)- i-	- zi(n)-	-ja -ja	animals	inja izinja	dog dogs
11 (10)	u- i-	(lu)- zi(n)-	-thi -thi	inanimates and long thin objects	uthi izinthi	stick sticks
14 15	u- u-	bu- ku-	-hle -cula	abstract nouns infinitives	ubuhle ukucula	beauty to sing
17		ku-		locatives, remote/ general		locative

Concordial agreement—example (isiZulu, South Africa)

Abafana abancane bazozithenga izincwadi ezinkulu
aba-fana **aba**-ncane **ba**- zo- **zi**- thenga **izi**-ncwadi e-**zi**-nkulu
2.boy **2**.small **2.SUBJ**-FUT-**10.OBJ**-buy **10**.book REL-**10**.big
'The little boys will buy the big books'

Other illustrative examples (isiZulu)

- 'and', enumerative: *na-*, phonologically conditioned
  - Ex: milk and butter: *ubisi <u>ne</u>bhotela* (-a+i-=-e-)
  - Ex: butter and milk: *ibhotela <u>no</u>bisi* (-a+u-=-o-)

Other illustrative examples (isiZulu)

- 'and', enumerative: *na-*, phonologically conditioned
  - Ex: milk and butter: *ubisi <u>ne</u>bhotela* (-a+i-=-e-)
  - Ex: butter and milk: *ibhotela <u>no</u>bisi* (-a+u-=-o-)
- 'is not a': combine NEG SC with PRON, both depend on noun class
  - Ex: an animal is not a plant: *isilwane <u>asiwona</u> umuthi*
  - Ex: a plant is not an animal: *umuthi <u>awusona</u> isilwane*

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*
- '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: *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

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[1]

[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[1]
- 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

$\Rightarrow$ 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

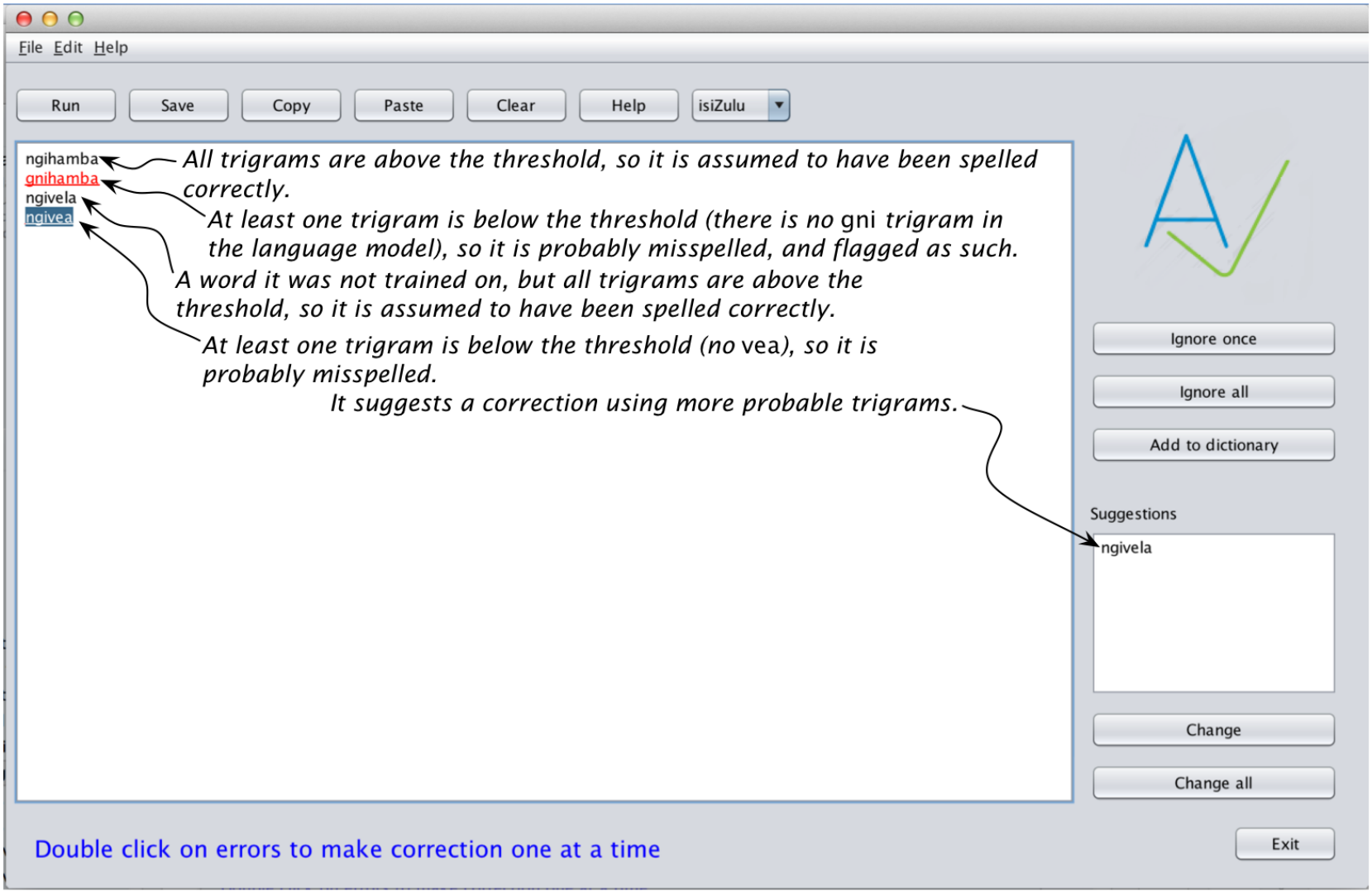
First iteration [Ndaba et al.(2016)]

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

corpus	generate trigrams	unique trigrams	frequency	probability
sivela	siv ive vel ela	ala	1	0.03703704
ngihamba	ngi gih iha ham amb mba	amb	1	0.03703704
uvelaphi	uve vel ela lap aph phi	aph	2	0.07407407
ngivala	ngi giv iva val ala	ela	3	0.11111111
uvelaphi	uve vel ela lap aph phi	gih	1	0.03703704
		giv	1	0.03703704
		ham	1	0.03703704
		iha	1	0.03703704
		iva	1	0.03703704
		ive	1	0.03703704
		lap	2	0.07407407
		mba	1	0.03703704
		ngi	2	0.07407407
		phi	2	0.07407407
		siv	1	0.03703704
		uve	2	0.07407407
		val	1	0.03703704
		vel	3	0.11111111

Example how this then works in the spellchecker

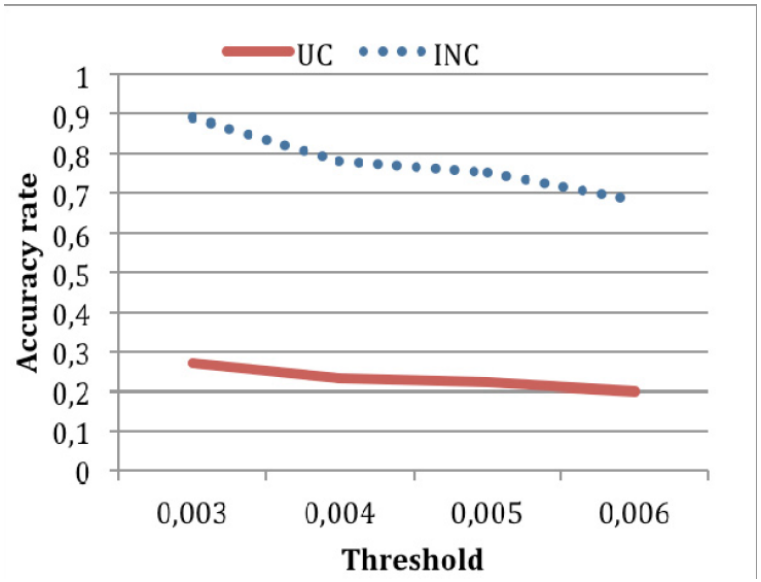


Basic approach for testing

- 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)

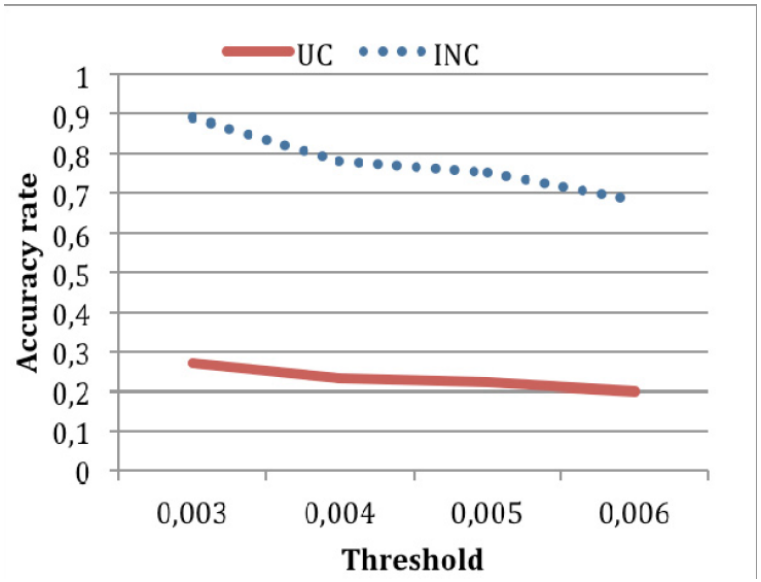
Major outcomes

- 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

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

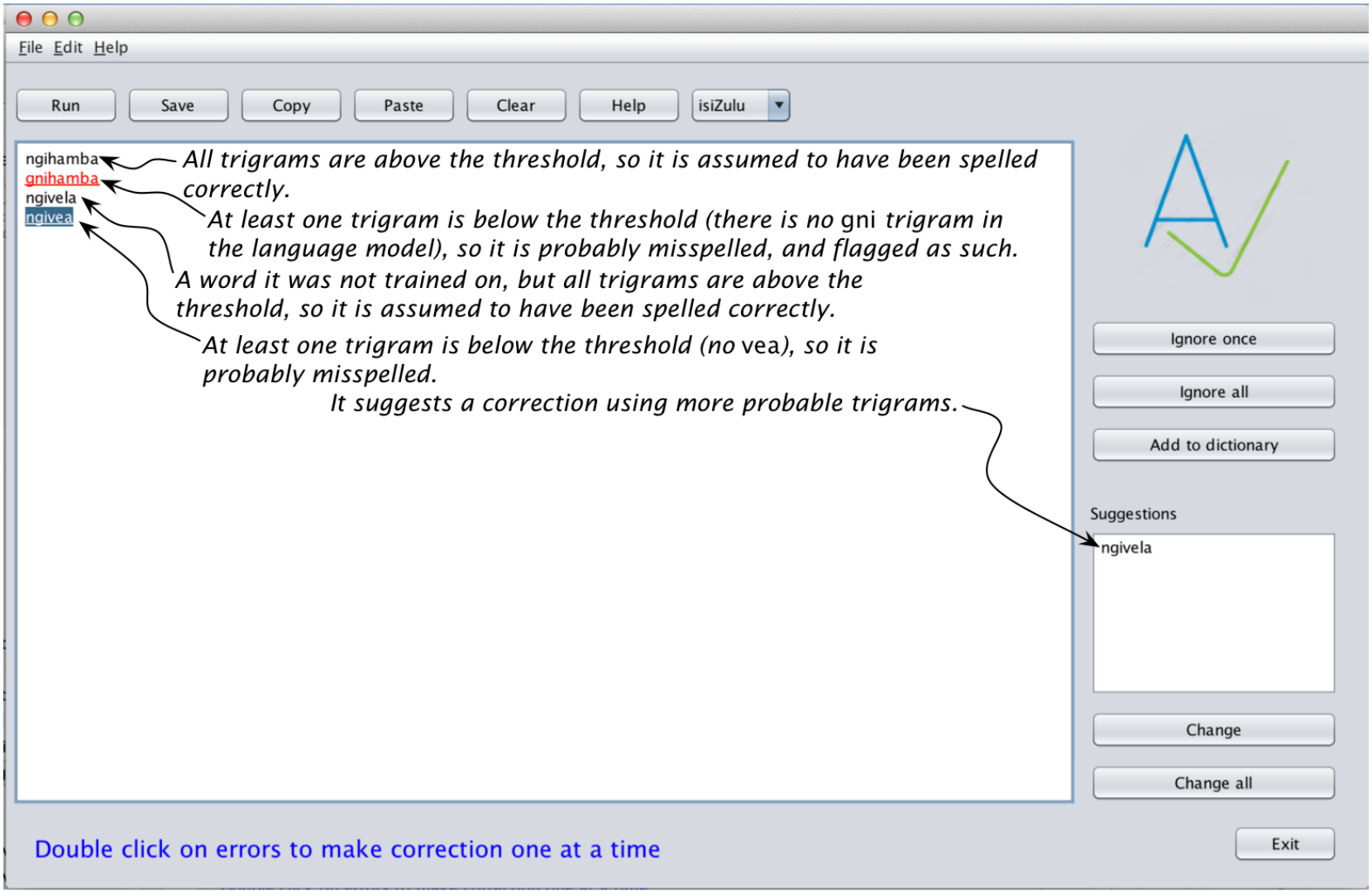
Error correction for isiZulu

- 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

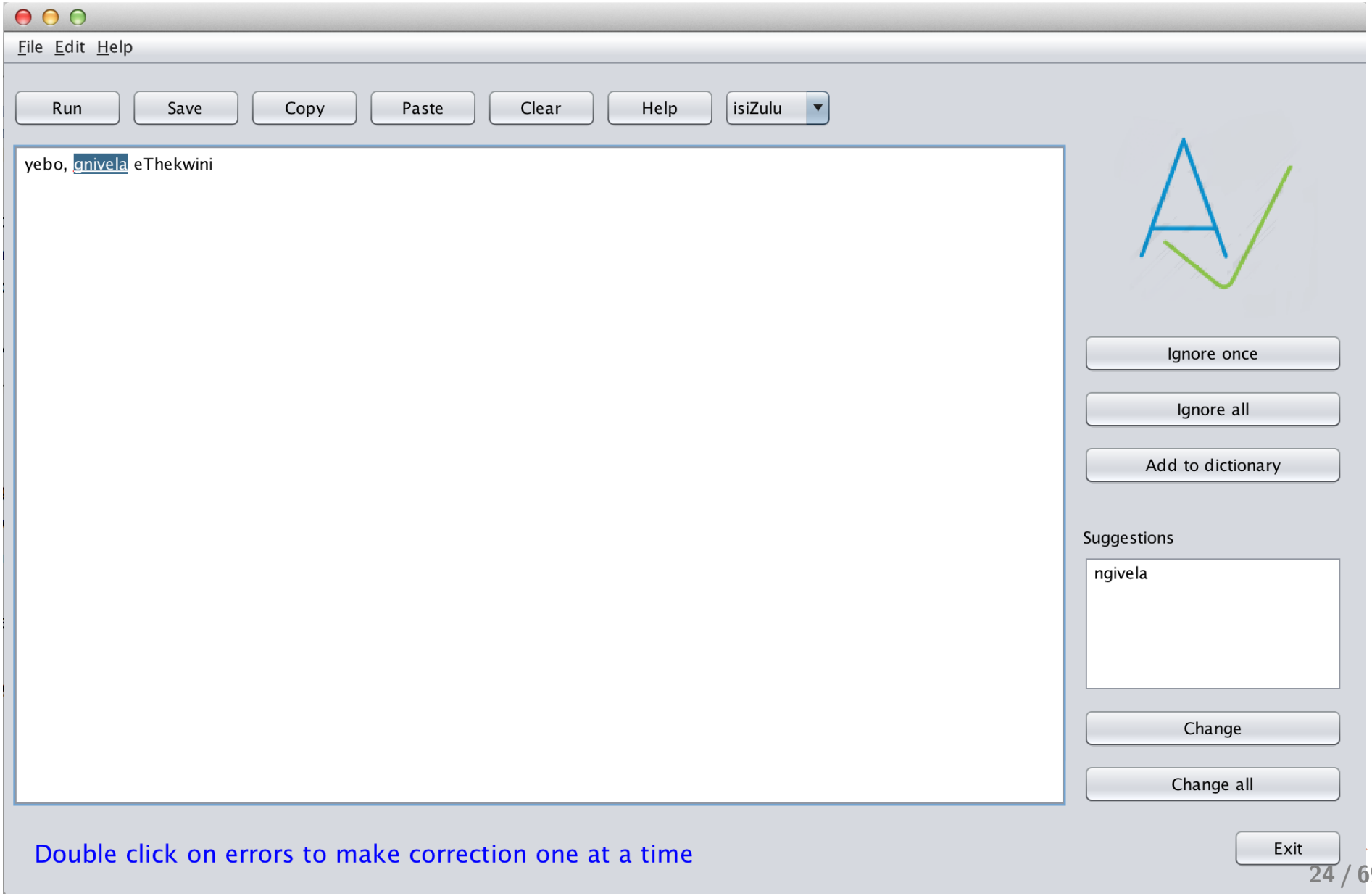
Error correction for isiZulu

- 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)

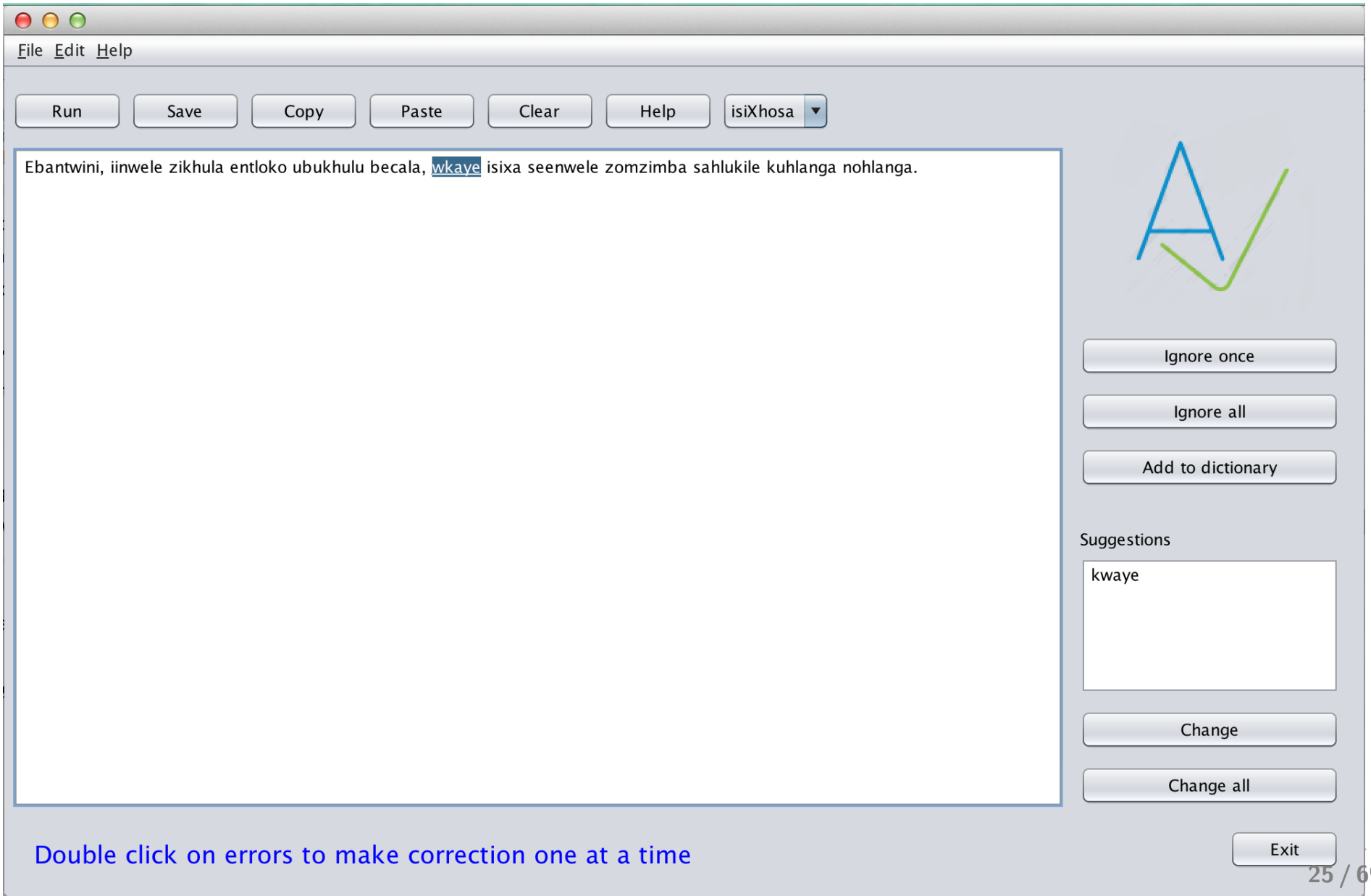
Deletion typo and a suggested correction (running example)



Error correction–transposition typo (isiZulu)



Error correction–transposition typo (isiXhosa)



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%

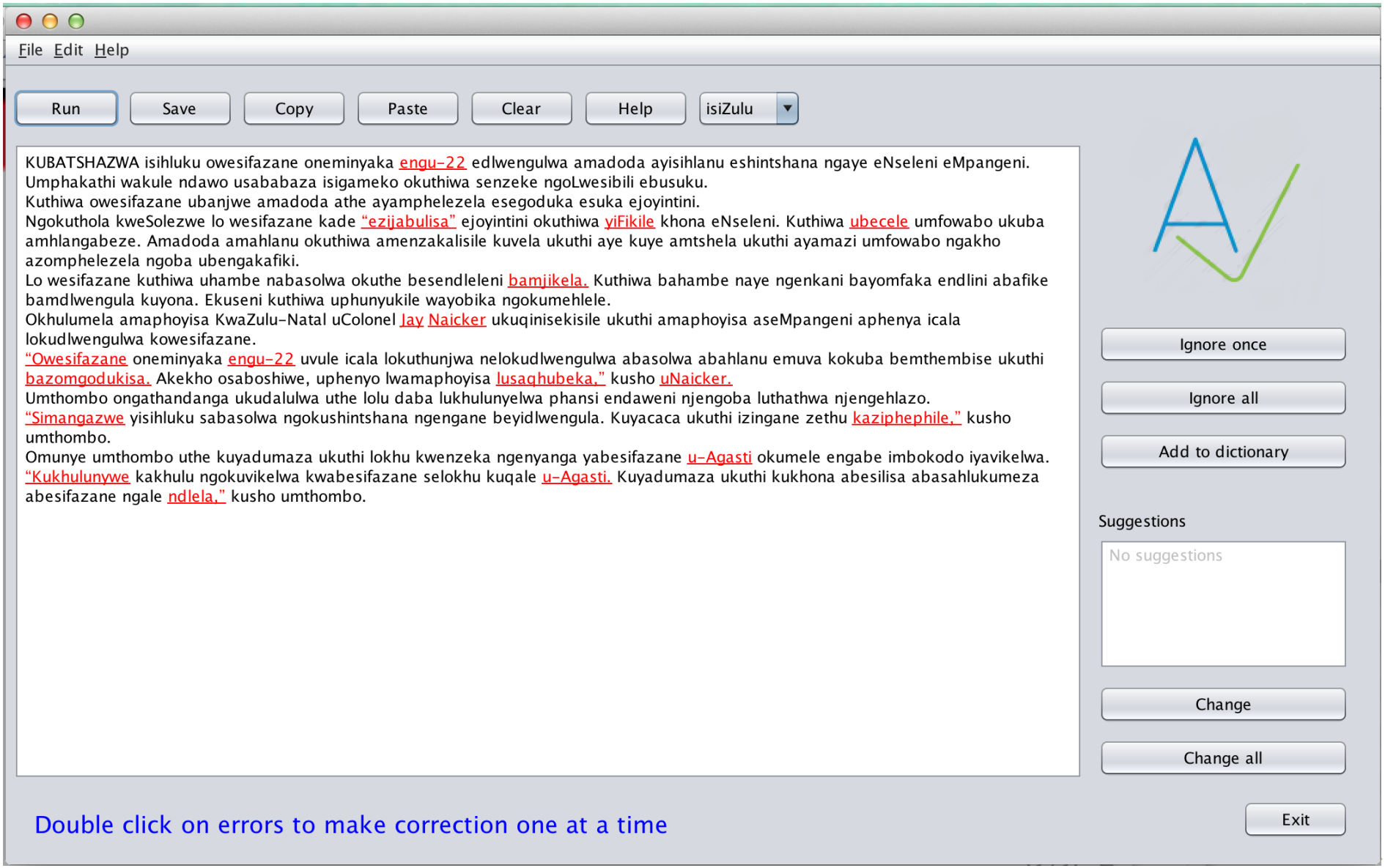
Results

- 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

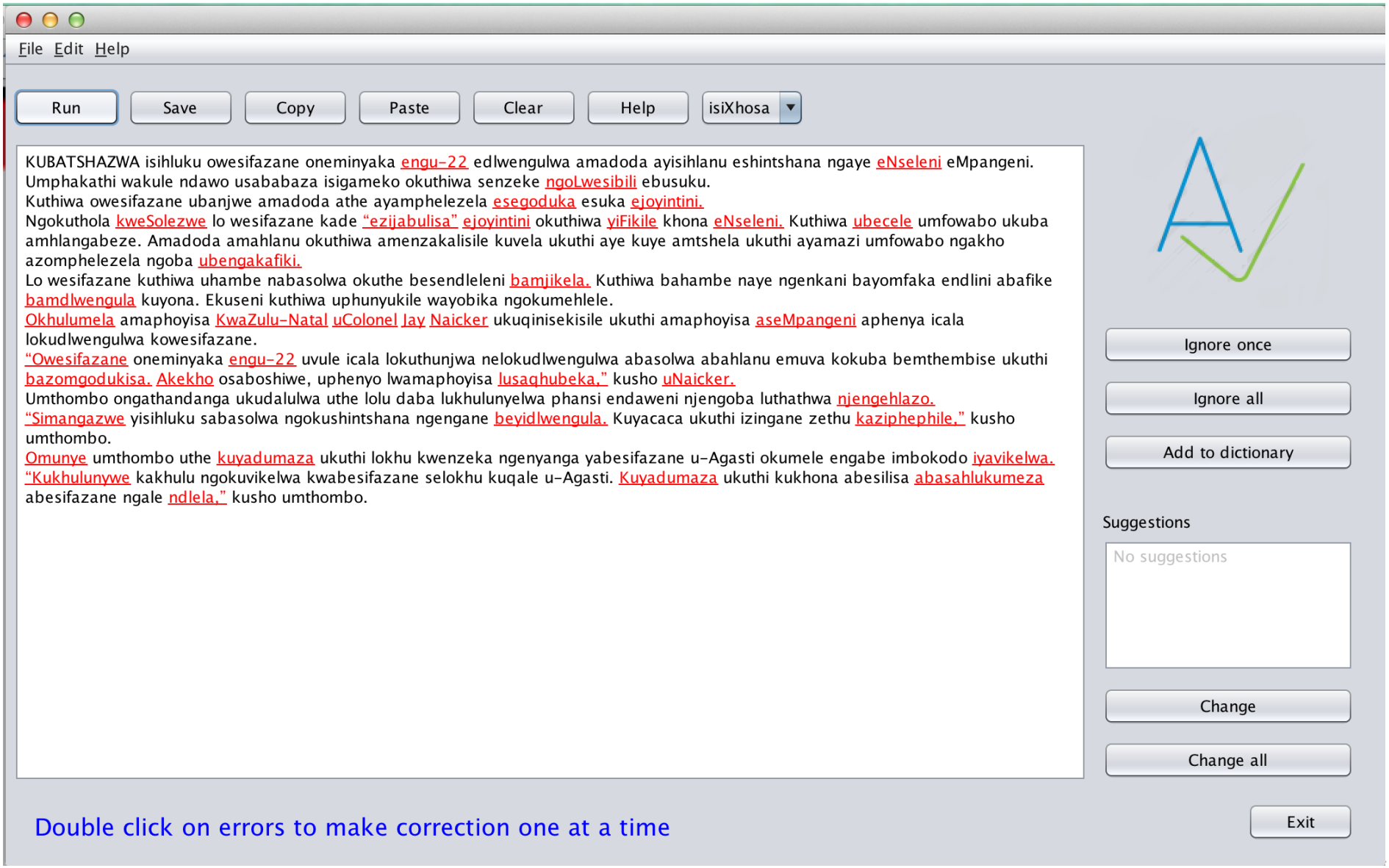
Spelling correction for isiXhosa

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

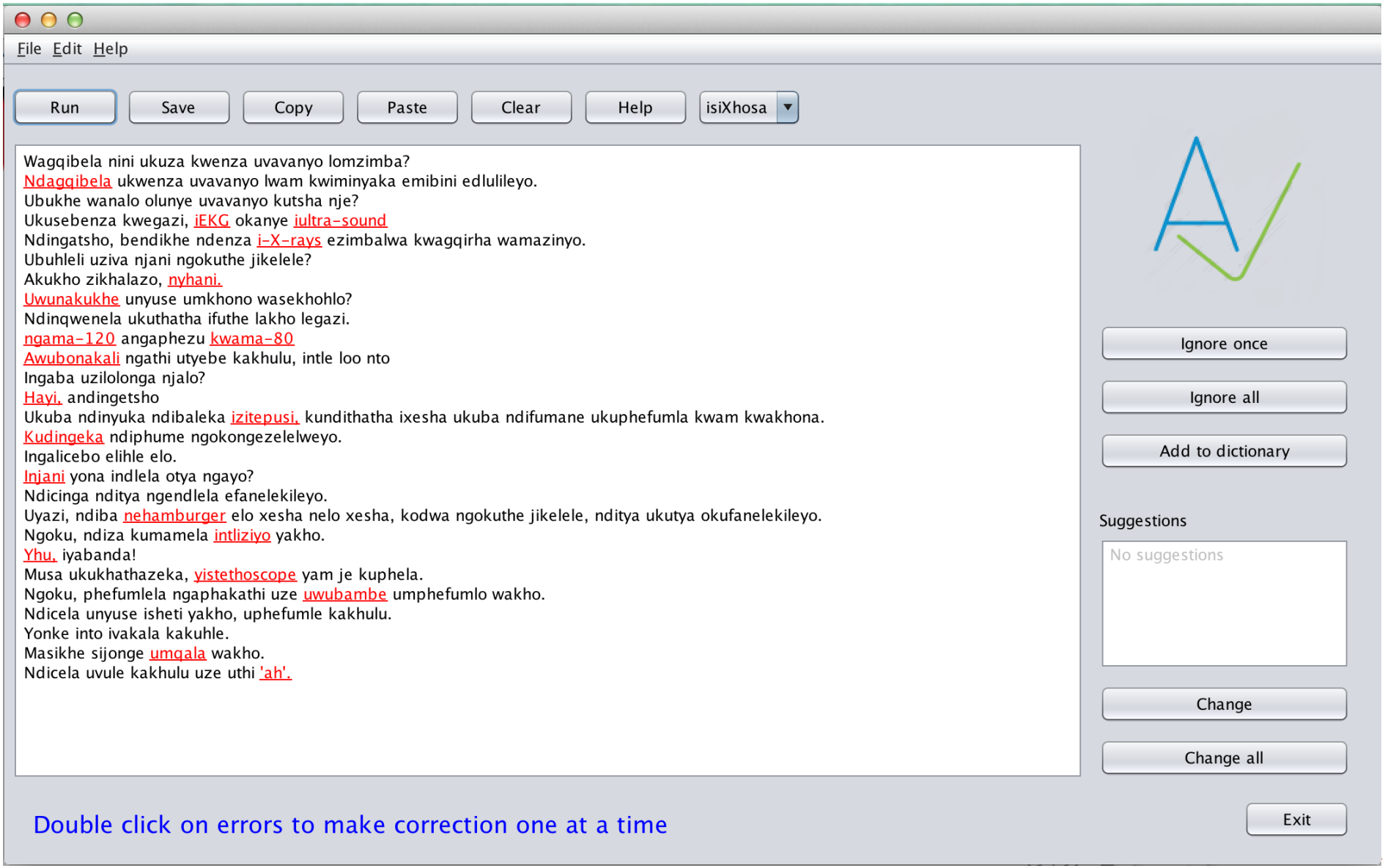
isiZulu text, isiZulu model



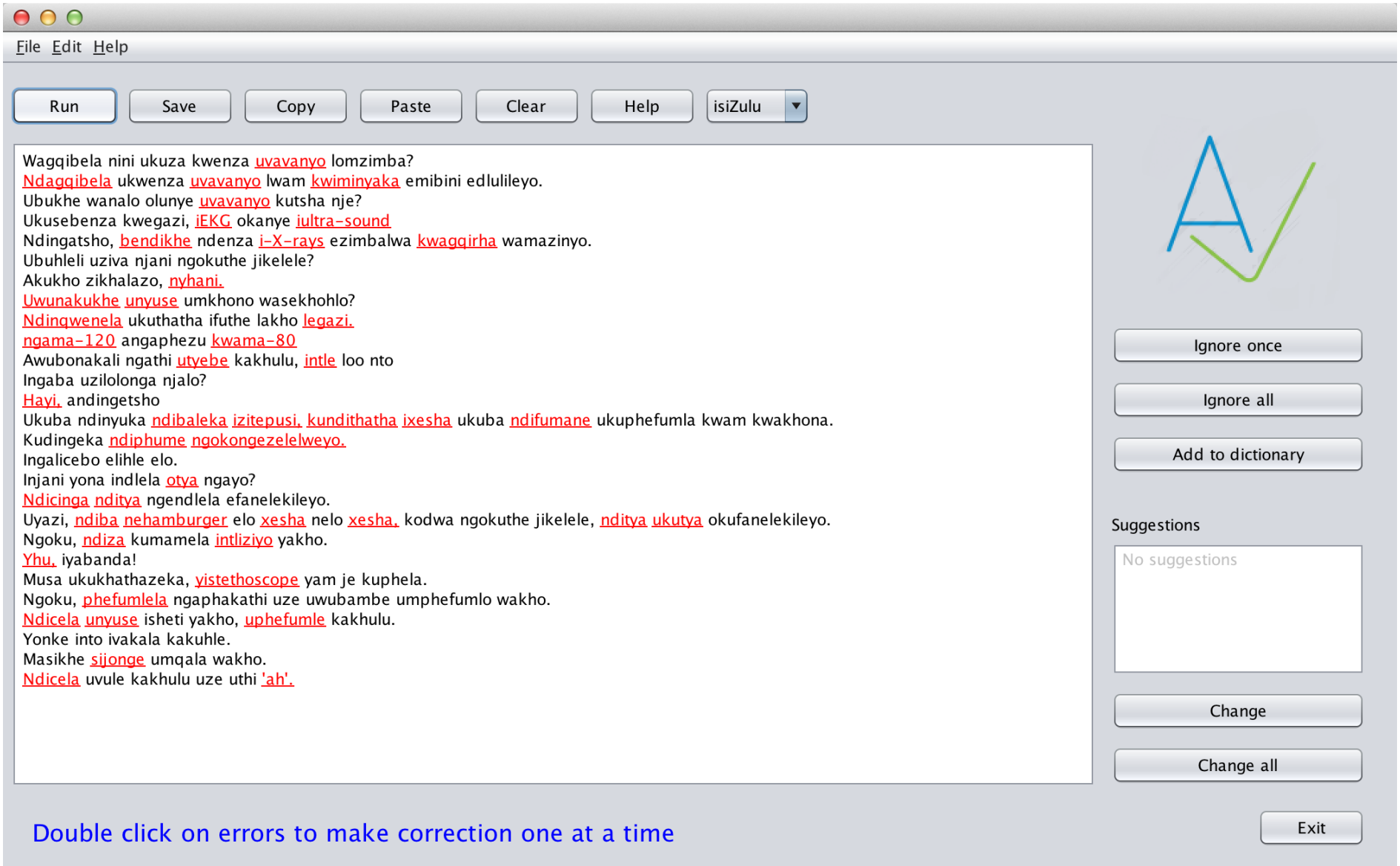
isiZulu text, isiXhosa model



isiXhosa text, isiXhosa model



isiXhosa text, isiZulu model



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

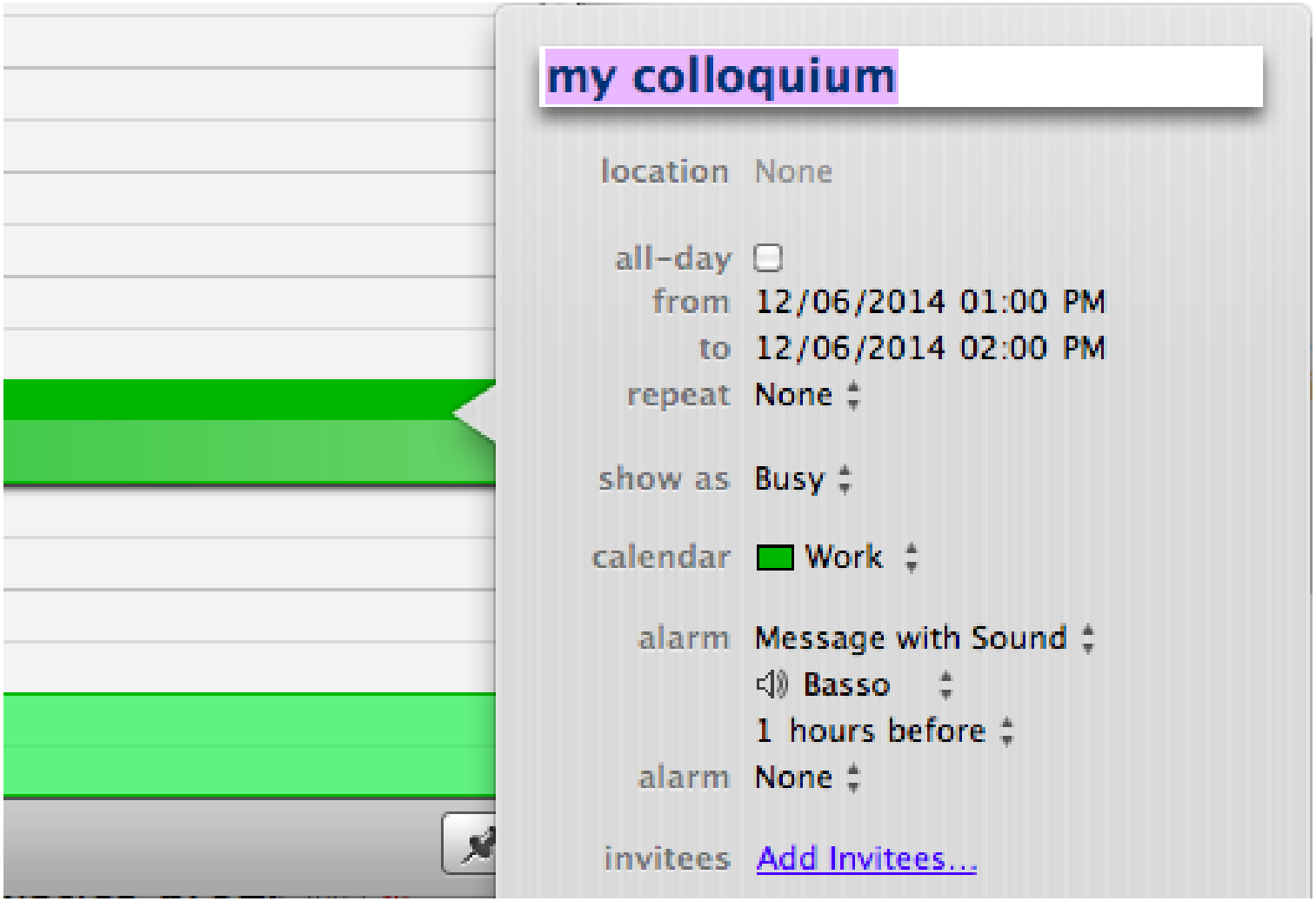
Short answer

- **C**controlled **N**aural **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

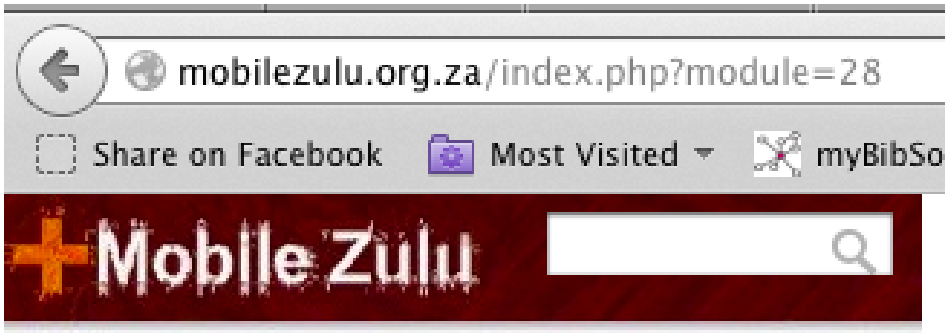
Natural language interfaces with some CNL or NLG

- 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.

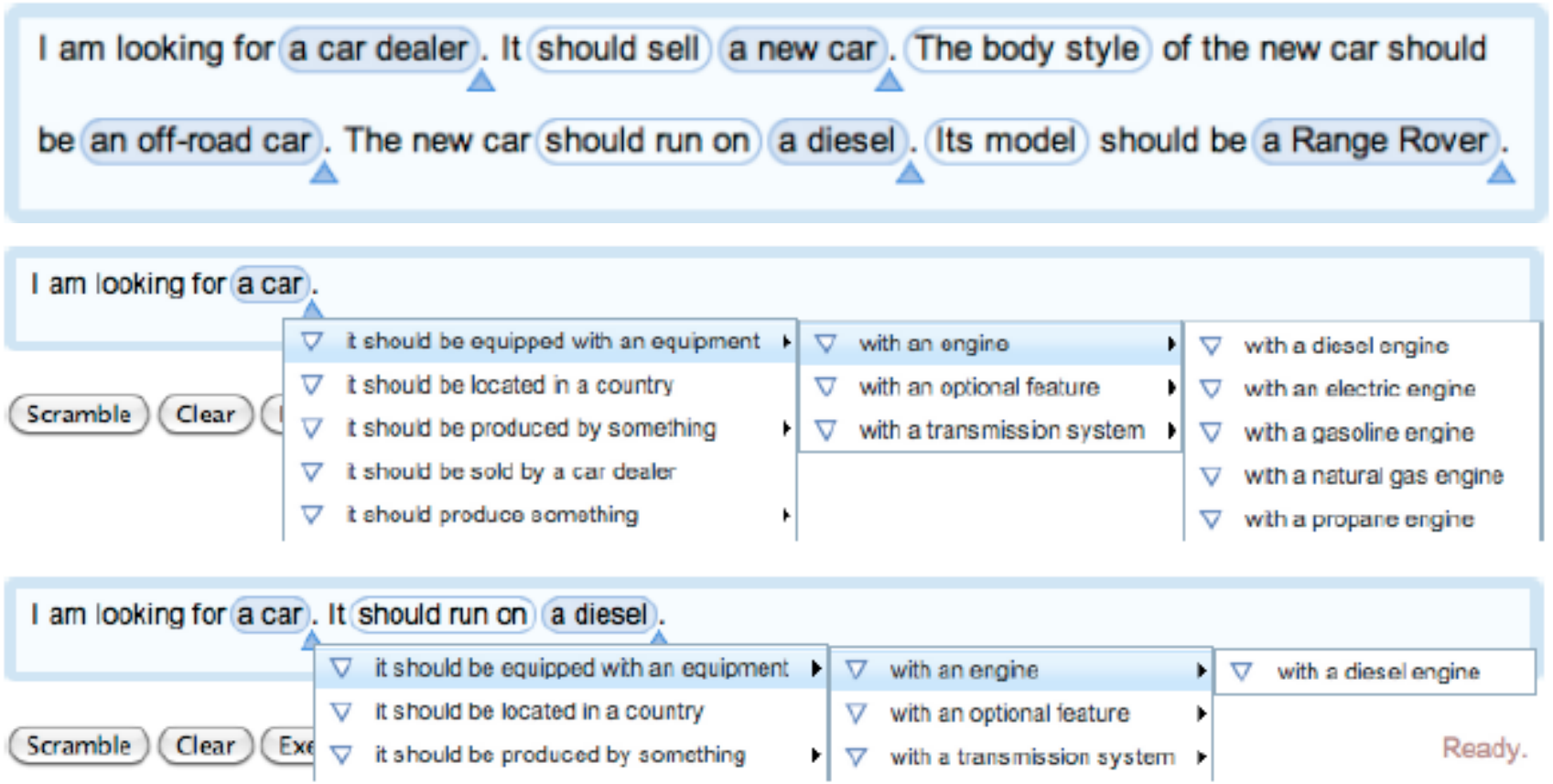
Example: iCal calendar entry with canned text



Example: Saadiq Moolla's mobile healthcare app



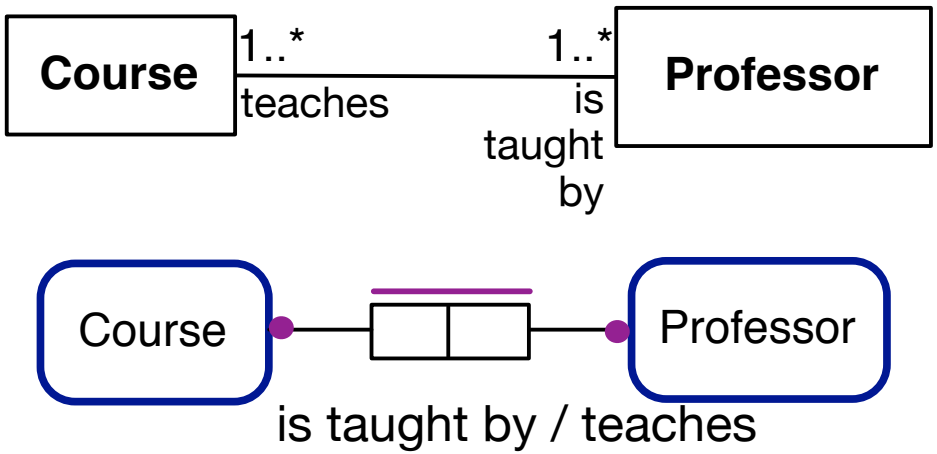
Example: Query formulation with Quelo [Franconi et al.(2010)]



Pictures from: Quelo @ The IESD Challenge 2012

Demo at: `http://krdbapp.inf.unibz.it:8080/quelo/`

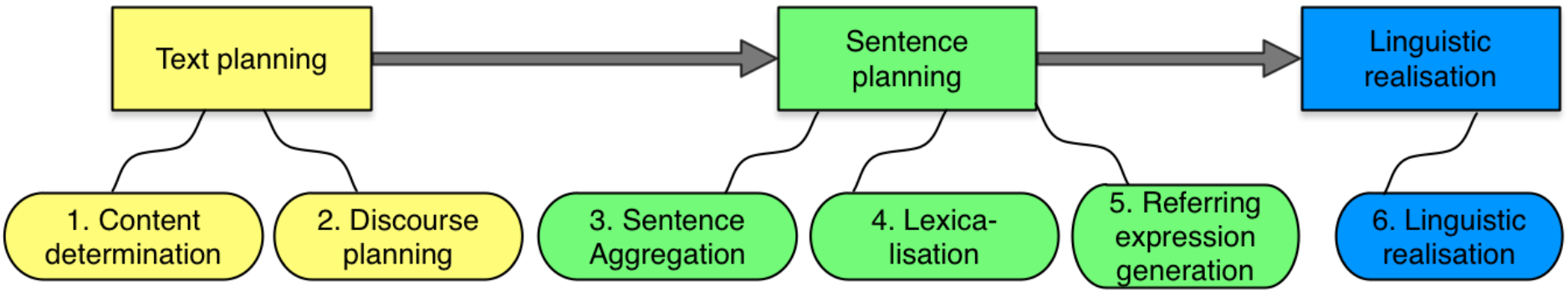
Example: Business rules and conceptual data models



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
  - Notably for English [Fuchs et al.(2010), Schwitter et al.(2008), Third et al.(2011), Curland and Halpin(2007)],
  - 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
  - Notably for English [Fuchs et al.(2010), Schwitter et al.(2008), Third et al.(2011), Curland and Halpin(2007)],
  - but also other languages [Jarrar et al.(2006)]
- Grammar engines, such as [Kuhn(2013)], Grammatical Framework (`http://www.grammaticalframework.org/`), SimpleNLG

$\Rightarrow$ CNL, NLG

Business rules/conceptual data models and logic reconstruction

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

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

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

Example of templates

```
<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

```
<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

```
<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

```
<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)]

NL Grammars, illustration

$$
\begin{aligned}
\textit{Sentence} &\longrightarrow \textit{NounPhrase} \mid \textit{VerbPhrase} \\
\textit{NounPhrase} &\longrightarrow \textit{Adjective} \mid \textit{NounPhrase} \\
\textit{NounPhrase} &\longrightarrow \textit{Noun} \\
&\ldots \\
\\
\textit{Noun} &\longrightarrow \textit{car} \mid \textit{train} \\
\textit{Adjective} &\longrightarrow \textit{big} \mid \textit{broken} \\
&\ldots
\end{aligned}
$$

(and complexity of the grammar)

Question

- 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)
- Details in [Keet and Khumalo(2014b), Keet and Khumalo(2014a), Keet and Khumalo(2017)]

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)

$\mathcal{ALC}$ syntax

- *Concepts* denoting entity types/classes/unary predicates/universals, including top $\top$ and bottom $\bot$;
- *Roles* denoting relationships/associations/n-ary predicates/properties;
- *Constructors*: and $\sqcap$, or $\sqcup$, and not $\neg$; quantifications forall $\forall$ and exists $\exists$
- *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
- *Individuals*

$\mathcal{ALC}$ semantics

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

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 $\sqsubseteq$

(U1) Boy $\sqsubseteq$ ...

wonke umfana ...	('each boy...'; *u-* + *-onke*)
bonke abafana ...	('all boys...'; *ba-* + *-onke*)

(U2) Phone $\sqsubseteq$ ...

lonke ifoni ...	('each phone...'; *li-* + *-onke*)
onke amafoni ...	('all phones...'; *a-* + *-onke*)

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

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

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

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` $\sqsubseteq$ `Plant`

ikhambi ngumuthi	('medicinal herb is a plant')
amakhambi yimithi	('medicinal herbs are plants')
wonke amakhambi ngumuthi	('all medicinal herbs are a plant')

(S2) `Giraffes` $\sqsubseteq$ `Animals`

izindlulamithi yizilwane	('giraffes are animals'; generic)

(S3) `Cellphone` $\sqsubseteq$ `Phone`

Umakhalekhukhwini uyifoni	('cellphone is a phone'; determ.)

Possible subsumption patterns

a. N_1 <copulative *ng*/*y* depending on first letter of N_2>N_2.

b. <plural of N_1> <copulative *ng*/*y* depending on first letter of plural of N_2><plural of N_2>.

c. <All-concord for NC_x>onke <plural of N_1, being of NC_x> <copulative *ng*/*y* 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*)

(SN1) $\texttt{Cup} \sqsubseteq \neg\texttt{Glass}$

indebe akuyona ingilazi ('cup not a glass')

zonke izindebe aziyona ingilazi ('all cups not a glass')

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

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

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

Possible negation (disjointness) patterns

a. <N_1 of NC_x> <NEG SC of NC_x><PRON of NC_y> <N_2 of NC_y>.

b. <All-concord for NC_x>onke <plural N_1, being of NC_x> <NEG SC of NC_x><PRON of NC_y> <N_2 with NC_y>.

Existential Quantification

(E1) Giraffe $\sqsubseteq \exists$eats.Twig

yonke indlulamithi idla ihlamvana elilodwa	('each giraffe eats at least one twig')
zonke izindlulamithi zidla ihlamvana elilodwa	('all giraffes eat at least one twig')

a. <All-concord for NC_x>onke <pl. N_1, is in NC_x> <conjugated verb> <N_2 of NC_y> <RC for NC_y><QC for NC_y>dwa.

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

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

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

Example

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

Example

- $\forall x\ (\text{uSolwazi}(x) \rightarrow \exists y\ (\text{ufundisa}(x, y) \wedge \text{Isifundo}(y)))$
- $\text{uSolwazi} \sqsubseteq \exists\ \text{ufundisa}.\text{Isifundo}$
- ?

$$\forall x\ (\text{uSolwazi}(x) \rightarrow \exists y\ (\text{ufundisa}(x, y) \wedge \text{Isifundo}(y)))$$

$$\text{uSolwazi} \sqsubseteq \exists\ \text{ufundisa}.\text{Isifundo}$$

$\forall x\ (uSolwazi(x) \rightarrow$ ... $x, y) \wedge Isifundo(y)))$

$uSolwazi \sqsubseteq \exists$ ufun...

look-up NC

pluralise

for-all

NC	AU	PRE
1	u-	m(u)-
2	a-	ba-
1a	u-	-
2a	o-	-
3a	u-	-
(2a)	o-	-
3	u-	m(u)-
4	i-	mi-
5	i-	(li)-
6	a-	ma-
7	i-	si-
8	i-	zi-
9a	i-	-
(6)	a-	ma-
9	i(n)-	-
10	i-	zi(n)-
11	u-	(lu)-
(10)	i-	zi(n)-
14	u-	bu-
15	u-	ku-
17		ku-

NC	QC (all) $QC_{oral+onke}$
1	u-onke → wonke
2	ba-onke → bonke
1a	u-onke → wonke
2a	ba-onke → bonke
3a	u-onke → wonke
(2a)	ba-onke → bonke
3	u-onke → wonke
4	i-onke → yonke
5	li-onke → lonke
6	a-onke → onke
7	si-onke → sonke
8	zi-onke → zonke
9a	i-onke → yonke
(6)	a-onke → onke
9	i-onke → yonke
10	zi-onke → zonke
11	lu-onke → lonke
(10)	zi-onke → zonke
14	ba-onke → bonke
15	ku-onke → konke

Bonke oSolwazi

$$\forall x\ (\text{uSolwazi}(x) \rightarrow \exists y\ (\text{ufundisa}(x, y) \wedge \text{Isifundo}(y)))$$

$$\text{uSolwazi} \sqsubseteq \exists \text{ufundisa}.\text{Isifundo}$$

... for relevant NC. Here:

AlgoConjugate

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

Bonke oSolwazi bafundisa

$$\forall x\ (\text{uSolwazi}(x) \rightarrow \exists y\ (\text{ufundisa}(x, y) \wedge \text{Isifundo}(y)))$$

$$\text{uSolwazi} \sqsubseteq \exists\ \text{ufundisa}.\text{Isifundo}$$

Bonke oSolwazi bafundisa Isifundo

$\forall x\ (\text{uSolwazi}(x) \rightarrow \exists y\ (\ldots) \wedge \ldots))$

uSolwazi $\sqsubseteq \exists$ ufundisa.I

look-up NC

get RC

get QC

add -dwa

NC	AU	PRE
1	u-	m(u)-
2	a-	ba-
1a	u-	-
2a	o-	-
3a	u-	-
(2a)	o-	-
3	u-	m(u)-
4	i-	mi-
5	i-	(li)-
6	a-	ma-
7	i-	si-
8	i-	zi-
9a	i-	-
(6)	a-	ma-
9	i(n)-	-
10	i-	zi(n)-
11	u-	(lu)-
(10)	i-	zi(n)-
14	u-	bu-
15	u-	ku-
17		ku-

RC	QC_{dwa}
o-	ye-
aba-	bo-
o-	ye-
aba-	bo-
o-	ye-
aba-	bo-
o-	wo-
e-	yo-
eli-	lo-
a-	wo-
esi-	so-
ezi	zo-
e-	yo-
a-	wo-
e-	yo-
ezi-	zo-
olu-	lo-
ezi-	zo-
obu-	bo-
oku-	zo-

Bonke oSolwazi bafundisa lsifundo esisodwa

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]

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

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

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.

M. Curland and T. Halpin.

Model driven development with NORMA.
In *Proceedings of the 40th International Conference on System Sciences (HICSS-40)*, pages 286a–286a. IEEE Computer Society, 2007.
Los Alamitos, Hawaii.

Enrico Franconi, Paolo Guagliardo, and Marco Trevisan.

An intelligent query interface based on ontology navigation.
In *Workshop on Visual Interfaces to the Social and Semantic Web (VISSW'10)*, 2010.
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.
Starlab technical report, Vrije Universiteit Brussel, Belgium, February 2006.
URL `http://www.meteck.org/files/ORMmultiverb_JKD.pdf`.

C. M. Keet and L. Khumalo.

Toward a knowledge-to-text controlled natural language of isiZulu.
Language Resources and Evaluation, 51(1):131–157, 2017.
doi: 10.1007/s10579-016-9340-0.

References II

C. Maria Keet and Langa Khumalo.

Toward verbalizing logical theories in isiZulu.
In B. Davis, T. Kuhn, and K. Kaljurand, editors, *Proceedings of the 4th Workshop on Controlled Natural Language (CNL'14)*, volume 8625 of *LNAI*, pages 78–89. Springer, 2014a.
20-22 August 2014, Galway, Ireland.

C. Maria Keet and Langa Khumalo.

Basics for a grammar engine to verbalize logical theories in isiZulu.
In A. Bikakis et al., editors, *Proceedings of the 8th International Web Rule Symposium (RuleML'14)*, volume 8620 of *LNCS*, pages 216–225. Springer, 2014b.
August 18-20, 2014, Prague, Czech Republic.

Langa Khumalo.

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

Tobias Kuhn.

A principled approach to grammars for controlled natural languages and predictive editors.
Journal of Logic, Language and Information, 22(1):33–70, 2013.

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.

References III

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.

R. Schwitter, K. Kaljurand, A. Cregan, C. Dolbear, and G. Hart.

A comparison of three controlled natural languages for OWL 1.1.
In *Proc. of OWLED 2008 DC*, 2008.
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.
In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING'10)*, pages 1020–1028.
Association for Computational Linguistics, 2010.
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 Unversity UK, 2011.
10th International Semantic Web Conference (ISWC'11), 23-27 Oct 2011, Bonn, Germany.

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