

Natural language generation for African languages

Zola Mahlaza

Department of Computer Science, University of Cape Town
zmahlaza@cs.uct.ac.za

Seminar @ UCT, March. 2023

Outline

1. Motivation
 - 1.1 Value of NLG systems
2. The past
3. The Present
 - 3.1 Patterns as a solution
 - 3.2 Improvements upon patterns
 - 3.3 Question generation
 - 3.4 Weather generation
4. The future
5. Why it matters?

Natural language and technology

1. Humans analyse, extract insights, and compile reports
 - ▶ South African Weather Service (SAWS): Numerical weather prediction data → weather forecast report
 - ▶ Financial institutions: financial data → report and slide decks
2. Challenges
 - ▶ Scalability and cost: \uparrow number of texts $\implies \uparrow$ cost
 - ▶ Issues relating to lang. ('multilingual')
3. Solution: computation. Data, information, knowledge → natural language text.
4. (Conversational) natural interfaces
 - ▶ Virtual assistants : Business Process Models → text
 - ▶ Etc.

Natural language generation

- ▶ Natural language generation \neq Machine Translation
- ▶ Natural language generation is a subfield Natural language processing

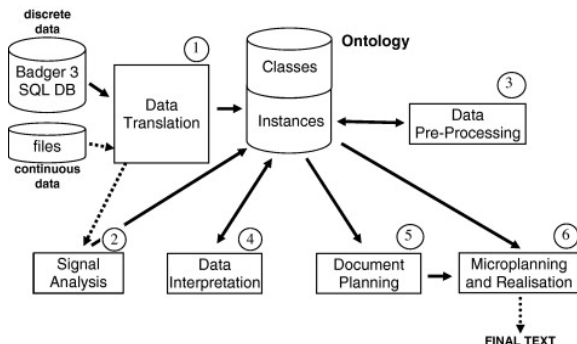


Figure: Architecture of the BT-Nurse system (Hunter et al. 2012)

Building natural language generating systems

- ▶ Strategy and tactics (Thompson 1977)
 - ▶ “What to say” and “How to say it”
- ▶ Three-step pipeline (Dale and Reiter 2000)
- ▶ End-to-end models (e.g., Castro Ferreira et al. 2019)

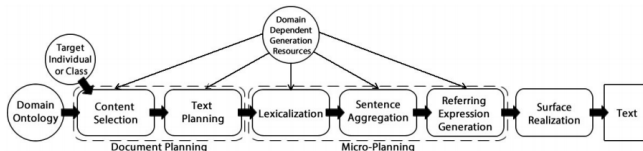


Figure: Knowledge-to-text system architecture used in NaturalOWL (Androutsopoulos et al. 2013)

Building natural language generating systems

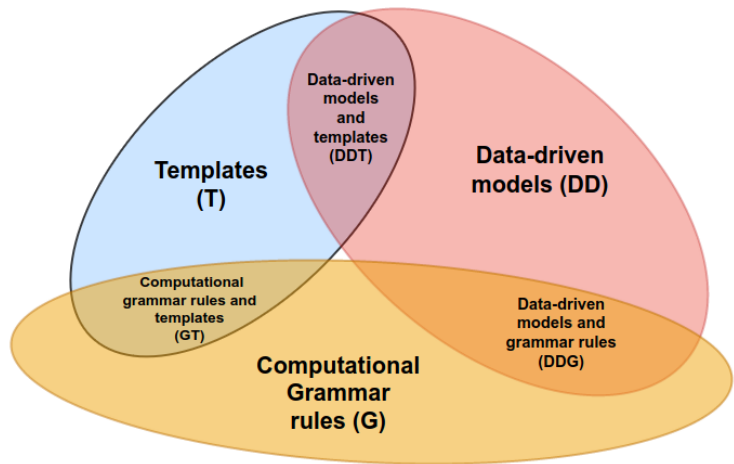


Figure: A classification of methods for generating text (Mahlaza, 2022)

Building natural language generating systems

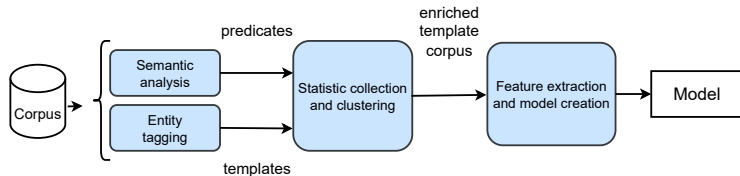
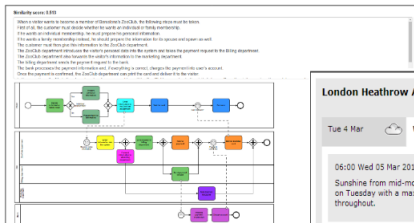
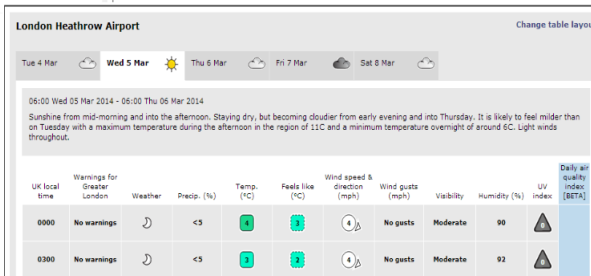


Figure: Representation of the process followed by (Howald et al. 2013)

Natural language and technology: limits



Screenshot of the BPMNvText module from the NLP4BPM suite (Delicado et al., 2017)



Screenshot of an automatically generated five-day weather forecast (Sripada et al., 2014)

Natural language and technology: limits

System/tool	Fam.	Input	Lang.
<i>Verbalizers</i>			
Davis et al. [98, 72]	C	GF	English, Dutch
Stevens et al. [264]	P	OWL	English
Kalparnad and Fuchs [124]	C	OWL	English
Lim and Halpin [160]	P	-	Malay, Mandarin
Andronopoulos, Loupouros, and Galanis [8]	EC	OWL	English, Greek
Grūnits, Nespere, and Soudite [100]	EC	OWL	Latvian
Davis et al. [73]	EP	OWL	English
Byanontasak, Keet, and DeRenzi [44]	EP	OWL	Burmese
Keet, Xakaza, and Khumalo [138]	EP	OWL	IsiZulu
Duanzilio [69]	C	OWL, GF	English, French, Italian, Finnish, Hebrew and Swedish
Hessain, Rajan, and Schwitter [112]	C	Rdf-ML, JSON	English
<i>NLG systems</i>			
Stenzhorn [263]	EP	XML	English, German, French, Italian, Russian, Bulgarian, Turkish
van Deuter, Thome, and Krahmer [273]	EP	-	English, Dutch, German
Wilcock [286]	EP	XML	-
<i>Surface realizers</i>			
McRoy, Channarukul, and Ali [190]	E	-	English
Bucmann [40]	E	Generation Interface Language	-

Name/Citation	Lang.
BioLcafflets [291]	English
Bio-AMR v0.8 [184]	English
SR18-ar [197]	Arabic
Rotowire [287]	English
SR18-pt [197]	Portuguese
SR18-fi [197]	Finnish
SR18-nl [197]	Dutch
SR18-cn [197]	Chinese
SR18-it [197]	Italian
SR18-es [197]	Spanish
SR18-fr [197]	French
WebNLG [92]	English
LogicNLG [57]	English
LDC2016E25 [184]	English
E2E [211]	English
SR18-rs [197]	Russian
SR18-cz [197]	Czech
ToTTo [219]	English
Percz-Beitrachini and Lapata [220]	English
Chisholm, Radford, and Hachey [58]	English
WikiBio [154]	English

Figure: Some of the prominent tools and datasets in NLG (Mahlaza, 2022)

Language diversity











Rank	Country	Total Languages	Population 2020 (M)
1	 Papua New Guinea	840	8.8
2	 Indonesia	711	270.6
3	 Nigeria	517	201.0
4	 India	456	1,366.0
5	 United States	328	328.2
6	 Australia	312	25.4
7	 China	309	1,398.0
8	 Mexico	292	127.6
9	 Cameroon	274	25.9
10	 Brazil	221	211.0

Figure: Top ten most linguistically diverse countries (World Economic Forum, 2021)

Motivation

- ▶ Maintaining linguistic diversity (e.g., Heritage and biodiversity, Creativity and innovation (Skutnabb-Kangas, 2002))
- ▶ Pure academic interest (e.g., methodological challenges posed by the other languages).
 - ▶ Generating text using a template.
 - ▶ "Hello **[name]**, please take a seat."

Methods have to be sensitive I

- ▶ Maties promotional material: "saam vorentoe · masiye phambili · forward together"
- ▶ Unnamed academic's website (last accessed 13 Feb 2023): "So let's Walk Together/Loop Saam/Hambani Kunye!"
 - ▶ Template = So let's **[translation]**

Methods have to be sensitive II

- ▶ Inadequacy of templates is known.
- ▶ Solution (circa 2015): use of patterns (Keet and Khumalo, 2014, 2017; Byamugisha et al., 2016)
 - Input** : indlovu $\sqsubseteq \exists$ idla.ihlamvana (i.e., elephant $\sqsubseteq \exists$ eats.twig)
 - Output** : *zonke izindlovu zidla ihlamvana elilodwa* 'all elephants eat at least one twig'
 - Pattern** : <QC(all) for NCx>onke <pl. N1 , is in NCx>
<conjugated verb> <N2 of NCy><RC for NCy> <QC for NCy>dwa;

Methods have to be sensitive III

Algorithm 2 Determine the verbalisation of existential quantification with object property (basic version, with conjugation)

1: \mathcal{C} set of classes, language \mathcal{L} with \sqsubseteq for subsumption and \exists for existential quantification;
variables: A axiom, NC_i noun class, $c_1, c_2 \in \mathcal{C}$, $o \in \mathcal{R}$, a_1 a term; r_2, q_2 concords; functions: $getFirstClass(A)$, $getSecondClass(A)$, $getNC(C)$, $getRC(NC_i)$, $getQC(NC_i)$, $getVSofOP(o)$.

Require: axiom A with a \sqsubseteq has been retrieved **and** an \exists on the rhs of the inclusion

```
2:  $c_1 \leftarrow getFirstClass(A)$  {get subclass}
3:  $c_2 \leftarrow getSecondClass(A)$  {get superclass}
4:  $o \leftarrow getObjectProp(A)$  {get object property}
5:  $v \leftarrow getVSofOP(o)$  {get verb stem of object property}
6:  $NC_1 \leftarrow getNC(c_1)$  {determine noun class by augment and prefix or dictionary}
7:  $NC_2 \leftarrow getNC(c_2)$  {determine noun class by augment and prefix or dictionary}
8:  $NC'_1 \leftarrow$  lookup plural nounclass of  $NC_1$  {from known list}
```

9: $c'_1 \leftarrow AlgoPluralize(c_1, NC_1)$

10: $a_1 \leftarrow$ lookup quantitative

11: $r_2 \leftarrow getRC(NC_2)$

12: $q_2 \leftarrow getQC(NC_2)$

13: **if** $checkNegation(A) ==$

14: {use Algorithm 3}

15: **else**

16: **if** o annotated with \exists

17: $conj_{nc1} \leftarrow$ lookup

18: $o' \leftarrow conj_{nc1} v$

19: $RESULT \leftarrow 'a_1 c'_1$

20: **else**

21: $RESULT \leftarrow 'passi$

22: **end if**

23: **end if**

24: **return** $RESULT$

Algorithm 3 Verbalisation of negation in an axiom (base cases: taxonomic subsumption and object property)

1: \mathcal{C} set of classes, language \mathcal{L} with \sqsubseteq for subsumption and \neg for negation; variables: A axiom, NC_i noun class, $c_1, c_2 \in \mathcal{C}$, a_1 term, a_2 letter and n, p are concords, v verb stem; functions: $checkNegation(A)$, $getNSC(NC_i)$, $getPNC(NC_i)$.

Require: $checkNegation(A) == true$

2: **select case**

3: negation directly preceded by \sqsubseteq **and** directly followed by c_2 **then**

4: $NC'_1 \leftarrow$ lookup plural nounclass of NC_1 {from known list}

5: $c'_1 \leftarrow AlgoPluralize(c_1, NC'_1)$ {call algorithm *AlgoPluralize* to generate a plural from o }

6: $a_1 \leftarrow$ lookup quantitative concord for NC'_1 {from quantitative concord (QC(all)) list}

7: $n \leftarrow getNSC(NC'_1)$ {get negative subject concord for c'_1 }

8: $p \leftarrow getPNC(NC_2)$ {get pronominal for c_2 }

9: $RESULT \leftarrow 'a_1 c'_1 np c_2.'$ {verbalise the disjointness (a_1 is QC(all))}

10: negation in front of OP **then**

11: $n \leftarrow getNSC(NC'_1)$ {get negative subject concord for c'_1 }

12: $RESULT \leftarrow 'a_1 c'_1 nvi c_2 r_2 q_2 dwa.'$ {verbalise the axiom}

13: negation in front of c_2 **and** A contains an OP **then**

14: $RESULT \leftarrow 'verbalisation of this class negation is not supported yet.'$

15: **end select case**

16: **return** $RESULT$

The problem with templates

1. Consider the scenario of a South African banking company
 - ▶ Customer base in region is largely Nguni-speaking members
 - ▶ Goal 1: customer visits to physical branches for certain matters.
 - ▶ Goal 2: encourage financial literacy via reports designed for behaviour modification
 - ▶ Decision: conversational agents and report generators¹ in Nguni languages
2. Limitations regarding re-usability and maintainability

¹e.g., https://projects.cs.uct.ac.za/honsproj/cgi-bin/view/2021/moraba_solomons.zip/

Recent developments I: the problem

There were no:

- ▶ approaches of pairing templates and grammar rules that prioritise the need to scaffold simple templates and reuse limited resource.
- ▶ no ontology-based specification of templates with support for morphologically rich languages
- ▶ architectures for creating an easy to maintain template-based surface realiser

hence, there are no Nguni language surface realisation tools that are easy to maintain and reusable.

Recent developments II: languages

1. IsiXhosa and isiZulu \in Niger-Congo B family. Largest in SA by L1 speakers.
2. Noun classes, agglutinating morphology, and concordial agreement.
3. Each noun belongs to 15-23 classes. Different classification systems
4. Example of a verb and agreement:

ba-sa-si-neth-isis-a

3pers pl-ASP_p-OC-rain_{VR}-INT-FV

'It is still raining intensely on us as a result of them'

Recent developments III: approaches and artefacts

- ▶ Develop a model-based approach to pairing templates and grammar rules (Mahlaza and Keet 2019, 2020).
- ▶ Created a task ontology for templates that support morphologically-rich languages (Mahlaza and Keet 2021)
- ▶ Develop an architecture to be used when organising surface realisation components for maintainable template-based realisers (Mahlaza and Keet, 2022).
- ▶ Created modular surface realisation engine for isiZulu and isiXhosa²
- ▶ Demonstrate the sufficiency of the developed approaches and artefacts for generating understandable and grammaticality correct isiZulu (Mahlaza and Keet 2020a) and isiXhosa text.

²<https://github.com/AdeebNqo/NguniTextGeneration>

Recent developments III: approaches and artefacts

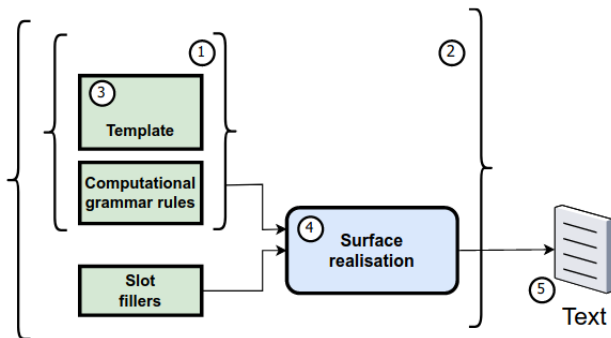


Figure: Relationship between the various elements (Mahlaza, 2022)

Recent developments IV: approaches and artefacts

- ▶ Develop a model-based approach to pairing templates and grammar rules (Mahlaza and Keet 2019, 2020).
- ▶ Created a task ontology for templates that support morphologically-rich languages (**Mahlaza2021**)
- ▶ Develop an architecture to be used when organising surface realisation components for maintainable template-based realisers (Mahlaza and Keet, 2022).
- ▶ Created modular surface realisation engine for isiZulu and isiXhosa³
- ▶ **Demonstrate the sufficiency of the developed approaches and artefacts for generating understandable and grammaticality correct isiZulu (Mahlaza and Keet 2020b) and isiXhosa text.**

³<https://github.com/AdeebNqo/NguniTextGeneration>

An isiZulu CNL for structured knowledge validation

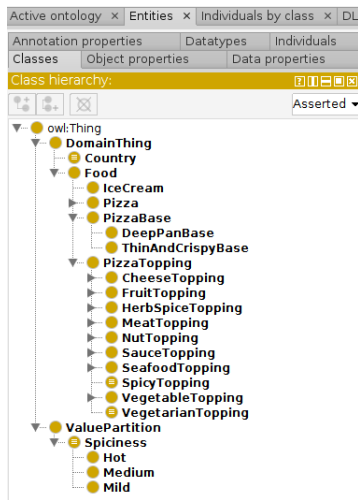


Figure: Screenshot of the pizza ontology

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment
- ▶ Step ≥ 2 : further increments + validate already codified knowledge

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment
- ▶ Step ≥ 2 : further increments + validate already codified knowledge
- ▶ Presenting the codified knowledge to experts

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment
- ▶ Step ≥ 2 : further increments + validate already codified knowledge
- ▶ Presenting the codified knowledge to experts
- ▶ Use controlled natural language (overview in (Softwat and Davis, 2017))

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment
- ▶ Step ≥ 2 : further increments + validate already codified knowledge
- ▶ Presenting the codified knowledge to experts
- ▶ Use controlled natural language (overview in (Saftwat and Davis, 2017))
- ▶ Observations, interviews, or task analysis based methods were already proposed in (Cooke 1994).
- ▶ Generate yes/no questions

An isiZulu CNL for structured knowledge validation

- ▶ Domain experts when building models or ontologies
- ▶ Step 1: Usable knowledge increment
- ▶ Step ≥ 2 : further increments + validate already codified knowledge
- ▶ Presenting the codified knowledge to experts
- ▶ Use controlled natural language (overview in (Saftwat and Davis, 2017))
- ▶ Observations, interviews, or task analysis based methods were already proposed in (Cooke 1994).
- ▶ Generate yes/no questions
- ▶ Language: isiZulu (L1 for 24% in South Africa)

Text generation from ontologies (1/2)

- ▶ Branches = educational question generators and model/ontology verbaliser
- ▶ Educational question generators:
 - ▶ English only
 - ▶ SimpleNLG (Gatt and Reiter, 2009) and/or regular templates
- ▶ Verbalisers:
 - ▶ IsiZulu, Runyankore, Afrikaans, English, Latvian, Mandarin, Bulgarian, Catalan, Danish, Dutch, Finnish, French, Hebrew, Italian, German, Norwegian, Romanian, Russian, Spanish, and Swedish.
 - ▶ Grammatical framework, basic templates, KPML, canned text, and grammar-infused templates

Text generation from ontologies (2/2)

- ▶ “Patterns” (Lim and Halpin 2016; Demey and Heath 2014; Keet and Khumalo 2017; Byamugisha et al. 2016)
- ▶ IsiZulu example from (Keet and Khumalo 2017):
 $QCall_{nc_x, pl} W_{nc_x, pl} SC_{nc_x, pl} - CONJ - P_{nc_y} RC_{nc_y} - QC_{nc_y} - dwa$
- ▶ Example output: “*Yonkeinja inekhanda elilodwa*” ‘Every dog has 1 head’
- ▶ Universal quantifier (“ $QCall_{nc_x, pl}$ ”) depends on noun “ $W_{nc_x, pl}$ ”
- ▶ Malay and Mandarin noun classifiers (Lim and Halpin 2016).
- ▶ IsiZulu and Runyankore noun dependencies (Keet and Khumalo 2017; Byamugisha et al. 2016)

OWL Simplified isiZulu (1/2)

- ▶ Select OWL axiom types from (Power 2012)
- ▶ Create templates for each axiom type
- ▶ Build verbaliser using Java
- ▶ Verbalised ontology from (Keet 2017). 91 axioms
- ▶ Internal validation by author
- ▶ External validation by isiZulu speakers: grammatically and understandability

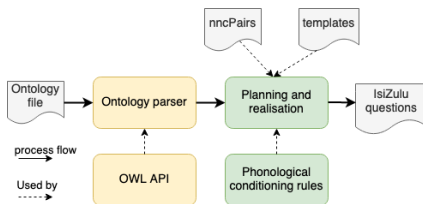


Figure: Verbaliser architecture

OWL Simplified isiZulu (2/2)

- ▶ SubClassOf, ClassAssertion, ObjectPropertyAssertion, EquivalentClasses, DisjointClasses, ObjectSomeValuesFrom, ObjectHasValue, DataPropertyAssertion, DataHasValue, ObjectAllValuesFrom, ObjectExactCardinality, ObjectMinCardinality, ObjectMaxCardinality
- ▶ 17 templates. Some axioms have multiple templates.
Example pairing:
SubClassOf(C1 C2)



Internal validation

- ▶ Verbalisable (76/91) and unverbalisable (15/91)
- ▶ Phonological conditioning errors (0/76)
- ▶ Morphological agreement errors (2/76)

DisjointClasses(isidlanyama isidlazitshalo)

(i) *asikho yini isidlanyama esiyisidlazitshalo?*

NEG-SC-exist carnivore_[NC7] RelC-COP-herbivore_[NC7]?

'Is there no carnivore that is a herbivore?'

DataPropertyAssertion(neminyaka uZola 50)

(ii) *Ingabe uZola neminyaka 50?*

Is Zola_[NC1a] CONJ-years 50?

'Is Zola aged 50?'

External validation (1/2)

- ▶ Six participants (five L1 isiZulu speakers and one L2)
- ▶ “grammatical and acceptable”, “grammatical and ambiguous”, “ungrammatical and understandable”, or “ungrammatical and unacceptable”

Table: Number of participants' judgements. Abbreviations: Pct. = percent

Survey	Gramm. + ambig.	Gramm. + ac- cept.	Ungramm. + under- stand.	Ungramm. + unac- cept.
A	17	41	6	12
B	23	78	19	32
A+B	40	119	25	44
A+B Pct.	18%	52%	11%	19%

External validation (2/2)

- ▶ Agree that texts 25 and 42 (template 3 and 4 respectively) ungrammatical and unacceptable

25 : *iNokia 3310 lifundisa uZola?*

'The Nokia 3310 teaches Zola'

42 : *Ingabe noma yiyiphi indlu evinyama?*

'Is every house (the same as) meat?'

Figure: Texts with errors underlined

- ▶ 83% of the texts positive. at most one participant judged 'ungrammatical and unacceptable'
- ▶ 71% of the texts positive. no participant judged 'ungrammatical and unacceptable'
- ▶ Disagreement in judgement not due to diff. in text length unlike (Keet and Khumalo 2014)
- ▶ Misunderstanding on text evaluation:
Ingabe lonke ibhotela lenza ifoni eliyi-1 ncamashi?
(‘Does every butter make exactly 1 phone?’)

Conclusions

- ▶ First isiZulu CNL and verbaliser generating questions for knowledge validation.
- ▶ Aggregated judgements by question, most questions (83%) are judged positively.
- ▶ When Survey A's criteria is relaxed, most questions (71%) are judged positively.
- ▶ Bad texts: noun class of 'phone' vs. 'nokia 3310', error in serialised template only

- ▶ English GALiWeather templates (Ramos-Soto et al. 2015)
- ▶ Textual short-term weather forecasts for every municipality in Galicia
- ▶ Example:
 - ▶ The temperatures will be [minT] for the minimums and [maxT] or the maximums compared to the expected for this time of the year , which globally will be [norV].
 - ▶ Iqondo eliphantsi lemozulu [minT] kwaye neqondo eliphezulu [maxT] xa lithelekiswa netempitsha elindelekileyo kwelixesha enyakeni, kodwa ndawo yonke itemprisha [norV]
- ▶ Captured the templates using the task ontology
- ▶ We evaluated 23 sentences (see (Mahlaza, 2022) for the generated text)
- ▶ Evaluation: fluency and grammaticality on a 5-point scale + (a single attention check question)
- ▶ Recruited participants via social media, encouraged participants to recruit other respondents

- ▶ 18 total responses (16 English, 2 isiXhosa instructions)
- ▶ All L1 isiXhosa speakers
- ▶ 2 failed the attention check (English instructions)
- ▶ 13/23 perceived as fluent and grammatically correct. No consensus on the rest
- ▶ Judging quality of texts without additional text for context (cf. selling ice-cream (Gkatzia et al. 2016)) and differences in dialects
 - ▶ "The temperatures will be low for this period of the year..."
 - ▶ /**ths**/ vs. /**th**/: *ndithi* 'I say' takes the form *ndithsi* (Nomlomo 1993)

Weather corpus and meaning

- ▶ Should a corpus be treated as gold standard? (Reiter and Sripada 2002)
 - ▶ Forecasters have different meanings for time terms (e.g., 'by evening')
 - ▶ Geographical referring expression generation (Ramos-Soto et al. 2016)

Weather corpus and meaning

- ▶ Daily Advisories from the South African Weather Service⁴
- ▶ Western Cape (02/March/2023): Cloudy with morning fog along the west-coast, otherwise fine and warm to hot but very hot over the central and eastern parts. The wind along the coast will be light to moderate westerly to south-westerly along the west-coast otherwise moderate to fresh easterly to south-easterly. The expected UVB sunburn index: Extreme

⁴[https:](https://www.weathersa.co.za/images/data/specialised/rsa_summ.pdf)

[//www.weathersa.co.za/images/data/specialised/rsa_summ.pdf](https://www.weathersa.co.za/images/data/specialised/rsa_summ.pdf)

How to proceed from here? What about other African languages?

What has been attempted?

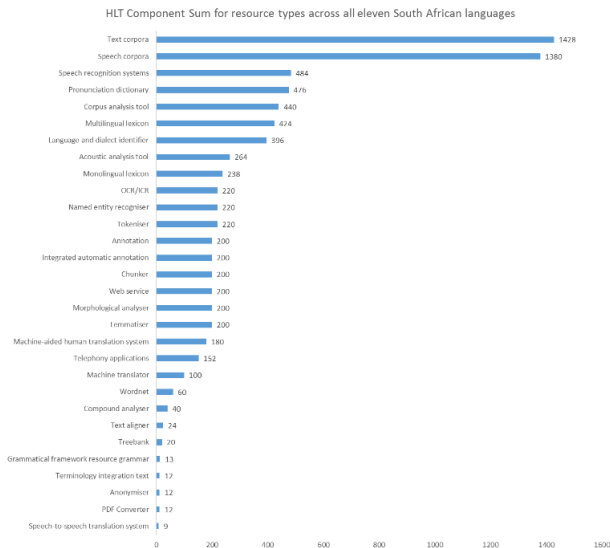


Figure: Human language technologies in South Africa (Wilken et al., 2018)

What has been attempted?

Language	Guthrie code	ISO 639-3	Task mentioned	Boot.
Chichewa	N31	nya	IR, similarity	-
Cinyanja	N31a?	nya	IR, similarity	-
Cisena	N44	seh	IR, similarity	-
Citonga	N15	tog	IR, similarity	-
Citumbuka	N21	tum	IR, similarity	-
Gikūyū	E51	kik	prefix extraction	+
Hunde	JD51	hke	-	-
Kaonde	L41	kqn	(computational cladistics)	-
Kimbundu	H21	kmb	similarity	-
Kinyarwanda	JD61	kin	noun class prediction	+
Kwangali	K33	kwn	(computational cladistics)	-
Luganda	JE15	lug	noun class prediction	+
isiNdebele (ZW)	S44	nde	morphological analysers	+
isiNdebele (ZA)	S408	nbl	pronunciation dictionary	+
isiXhosa	S41	xho	NLG, MT, morphological analysers, pronunciation dictionary, similarity	+
isiZulu	S42	zul	NLG, morphological analysers, MT, prefix extraction, POS tagger, corpus development, spellchecker, pronunciation dictionary, similarity	+
Mboshi	C25	mdw	-	-
Mpiemo	A86c	mcx	POS tagger	+
Ngoni	N12	ngo	morphological analysers	+
Pokomo	E71	pkb	(computational cladistics)	-
Runyankore	JE13	nyn	NLG, similarity, (computational cladistics), noun class prediction	+
Sanga	L35	sng	(computational cladistics)	-
Sepedi	S32	nso	POS tagger, pronunciation dictionary, similarity	+
Setswana	S31	tsn	morphological analysers, pronunciation dictionary, similarity	+
Shona	S10 (S11-15)	sna (twl, mxc, twx, ndc)	MT, similarity	+
siSwati	S43	ssw	morphological analysers	+
Swahili	G40 (G41-43)	swa, swl (ccl, sta)	MT, POS tagging, pronunciation dictionary, news item monitoring, similarity, (computational cladistics)	+
Swahili (Congolese)	G40g	swc	MT	+

Figure: List of NLP tools for African languages and employment of bootstrapping strategies (Keet, 2022)

How to proceed?

- ▶ 'Low-level' tasks (e.g., noun class prediction/disambiguation, etc.)
- ▶ Massively multilingual (e.g., language identification for 517 African languages (Adebara et al. 2022))
- ▶ Large number of heterogeneous datasets
- ▶ Opportunities for data integration or federation
- ▶ How well do your models generalize across datasets?
- ▶ Data augmentation and bootstrapping in a multilingual setting

Why focus on these tasks?

Last Words

Boring Problems Are Sometimes the Most Interesting

Richard Sproat
Search Google, Japan
rws@google.com

In a recent position paper, Turing Award Winners Yoshua Bengio, Geoffrey Hinton, and Yann LeCun make the case that symbolic methods are not needed in AI and that, while there are still many issues to be resolved, AI will be solved using purely neural methods. In this piece I issue a challenge: Demonstrate that a purely neural approach to the problem of text normalization is possible. Various groups have tried, but so far nobody has eliminated the problem of unrecoverable errors, errors where, due to insufficient training data or faulty generalization, the system substitutes some other reading for the correct one. Solutions have been proposed that involve a marriage of traditional finite-state methods with neural models, but thus far nobody has shown that the problem can be solved using neural methods alone. Though text normalization is hardly

Why focus on these tasks?



Figure: <https://lelapa.ai/comparing-africa-centric-models-to-openais-gpt3-5-2/>

Why focus on these tasks?

yintoniinja?

Inja (Canis lupus familiaris) iyiphi indidi yezilwanyana eziphuma kwimvelo yakwa-Afrika. Inja ikhetha ukuba amabhunga amacala, abanye basebenzisa le nto ukuba baphethe amahashe. I-Canidae udidi lwezilwanyana luqinisekisa ukubainja iyinto yokulungelelanisa, yokufumana iintlungu, yokusungula isigqwesha kwaye yokubaleka kwizimvo zentlalo. Inja iyinto enhle kakhulu kwizinto ezizizalayo, ezibonakalayo, ezisemgangathweni, ezizalisayo, ezisemgangathweni kunye nezinye iinkanuko. Inja ikhetha ukuba ibe namabhunga amacala ezifana nokubulala amaninzi, okungenakwenzeka lokho kungenxa yokuba iyinto edliwayo, edliwayo kunye nokulahleka kuluntu. Inja ikhetha ukuba ibe namabhunga amacala ezifana nokubulala izinja ezingenakwenzeka, ngokuba iimpawu zangaphakathi zazo zingena kwimvelo yonke imihla.

Masibuyel'embo - iingcambu zesiXhosa - Inja sisilwanyana ...

<https://www.facebook.com/1640875339510225/photos/inja-sisilwanyana-sasekhaya1yintoni-inkolelo-esinayo-...>

Inja - Wikipedia

|



Figure: Output of YouChat to the question 'what is a dog' in isiXhosa?

Why focus on these tasks?

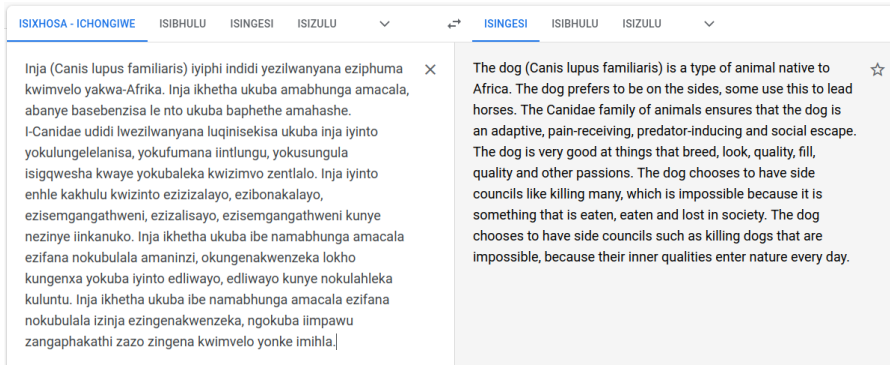


Figure: Automatic translation of the output from YouChat to the question 'what is a dog' in isiXhosa?

Last word

- ▶ Hons. students: consider doing a masters degree!
- ▶ My contact details: Office 3.06.2, zmahlaza@cs.uct.ac.za