GF application grammars in a resource scarce environment

Laurette Marais

CSIR
Touching lives through innovation

Outline

What is resource scarce?

Why use GF application grammars?

AwezaMed S2S translation app

Which corners to cut?

NN instead of concrete syntax

    Question-understanding

Conclusion

What is resource scarce?

Computational and linguistic skills

- Education in South Africa
  - ~ 1.1M Grade 1's in 2008, ~ 441K passed senior certificate in 2020 (with 40+%)
  - Senior certificate is written in Afrikaans or English
  - All languages can be written as home language and first or second additional language
  - Zulu HL: 143 364, Ndebele HL: 4 621
  - By Grade 5, 63% of learners have not acquired basic mathematical knowledge[1]
  - Mathematics 60+%: 30 882

1. *Trends in International Mathematics and Science Study 2019*, https://timssandpirls.bc.edu/timss2019

What is resource scarce?

Zulu

- 11.5M HL speakers (2011)
- Official language of South Africa (one of 11)
- 9 174 Wikipedia articles (seems to be mostly stubs), 28 active registered users
- Largest parallel English-Zulu corpora:
  - JW300 ~1.1M sentences
  - Autshumato ~35k sentences
- Morphological analyser, RG in progress[2]

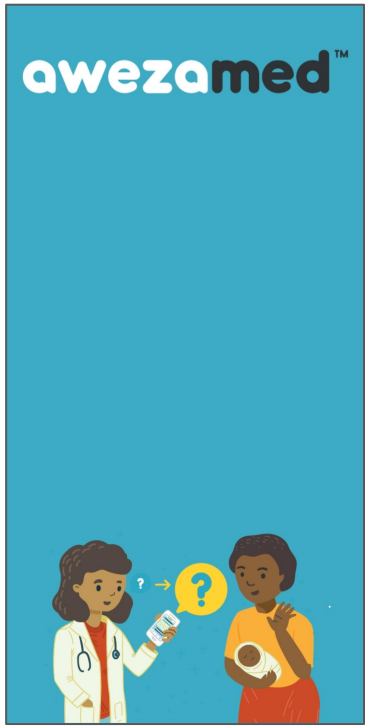
2. *https://github.com/LauretteM/gf-rgl-zul*

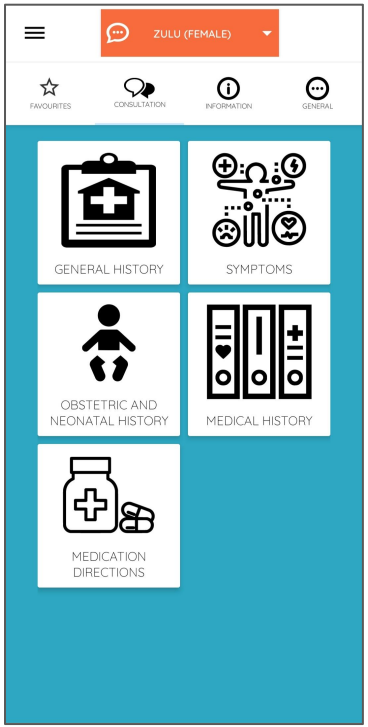
Why use GF application grammars?

The voice applications we target tend to

- Be high risk (why?)
- Involve “producer tasks”
- Concern a limited domain
- Be required in multiple (resource scarce) languages

AwezaMed S2S translation app





Maternal health app that translates ~4000 English utterances to Afrikaans, Xhosa and Zulu

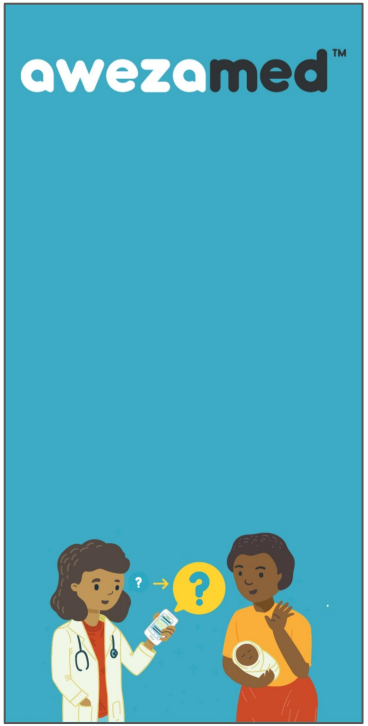
Supported by 5 multilingual (multimodal) GF applications grammars and mini resource grammars

AwezaMed S2S translation app

Enables communication during a consultation between

- English-speaking HCPs such as doctors, sisters and nurses
- non-English speaking patients

Uses ASR, TTS and GF MT to achieve S2S translation



AwezaMed S2S translation app

Maternal health domain (midwifery and obstetrics)

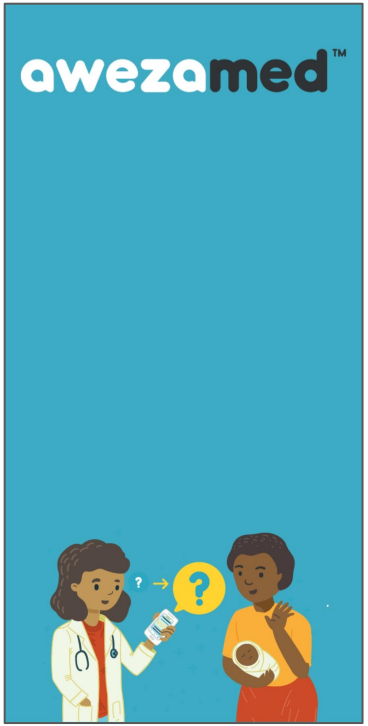
In South Africa, practised at

- Clinics
- Community Health Centres
- Hospitals

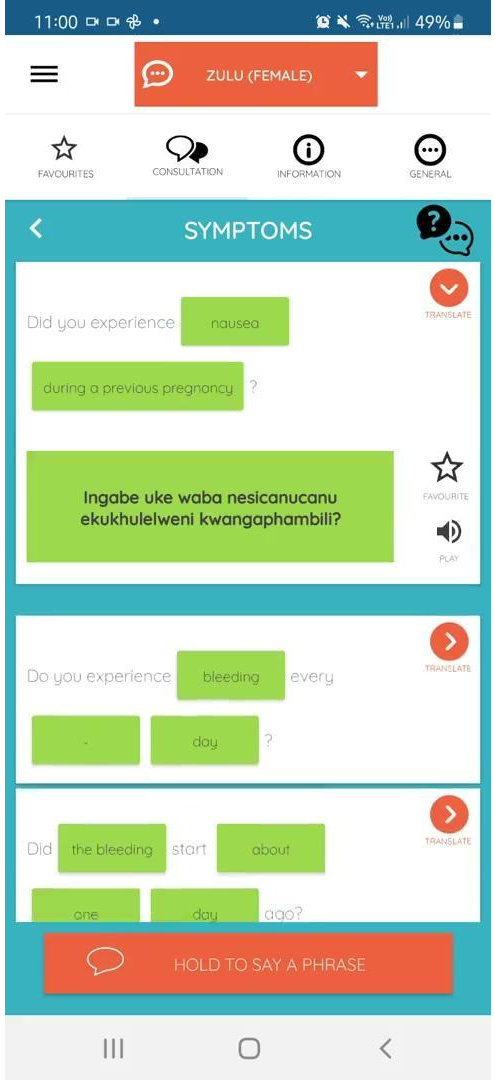
HCP responsible for

- respectful care → reliable, verifiable translation

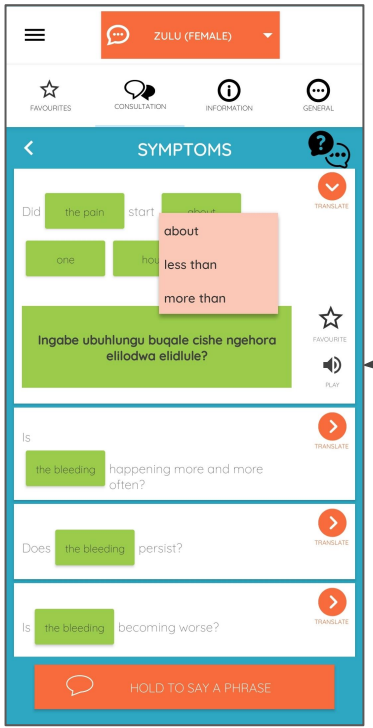
App uses questions requiring “yes” or “no”



Demo


11:00
49%
ZULU (FEMALE)
FAVOURITES
CONSULTATION
INFORMATION
GENERAL
SYMPTOMS
TRANSLATE
Did you experience
nausea
during a previous pregnancy
?
Ingabe uke waba nesicanucanu ekukhulelweni kwangaphambili?
FAVOURITE
PLAY
TRANSLATE
Do you experience
bleeding
every
day
?
TRANSLATE
Did
the bleeding
start
about
one
day
ago?
HOLD TO SAY A PHRASE

AwezaMed S2S translation app

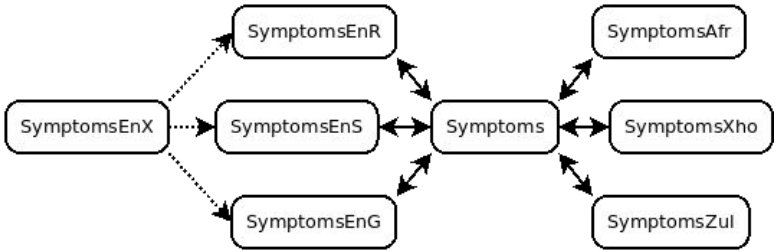


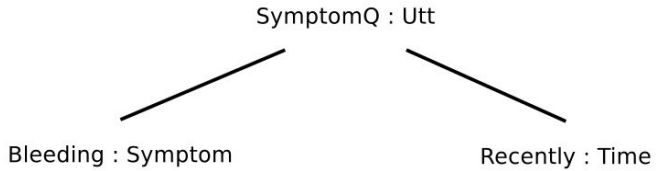
“did the pain start about two hours ago”

Did [Symptom: the pain] start [PointTime: about] [SmallNumber: two] [TimeMeasure: hours] ago?

“Ingabe ubuhlungu buqale ngaphezu kwehora elilodwa elidlule?”

AwezaMed S2S translation app





English

Have you experienced bleeding recently?

Have you experienced severe headaches recently?

Zulu

Ingabe uke waba nokopha kungekudala?

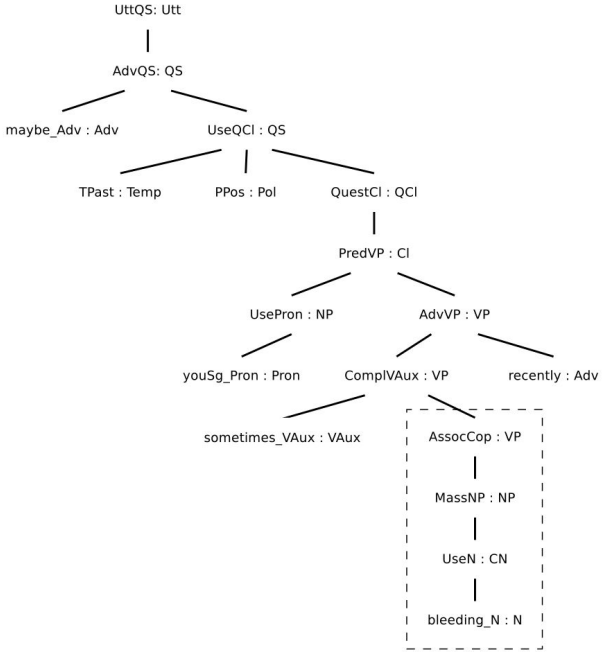
Maybe you were sometimes with bleeding recently?

Ingabe uke waphatwa yikhanda elibuhlungu kungekudala?

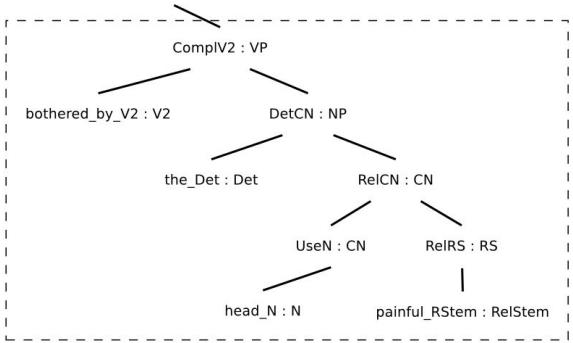
Maybe sometimes you were bothered by a head that is painful recently?

AwezaMed S2S translation app

“Ingabe uke **waba** **nokopha** kungekudala?”



“Ingabe uke **waphatwa** **yikhanda** **elibuhlungu** kungekudala?”



AwezaMed S2S translation app

Piloted at 5 sites in 3 provinces of South Africa

- Gauteng
- Western Cape
- KwaZulu-Natal

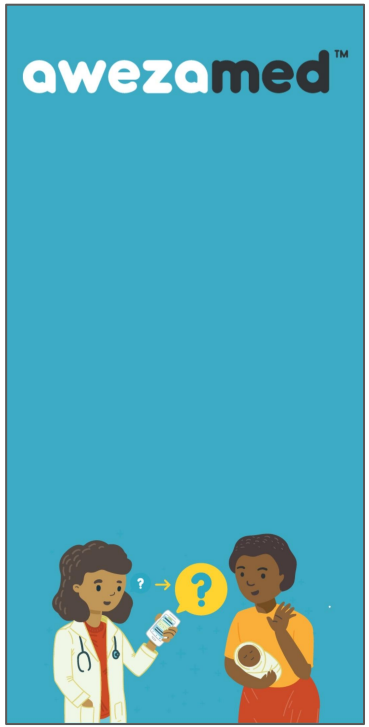
Feedback

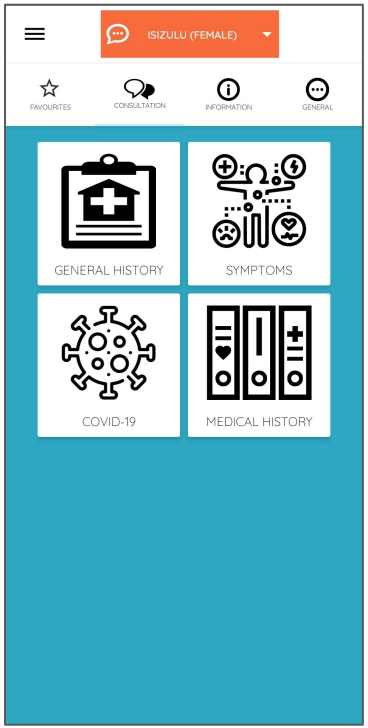
- Useful, especially in rural areas
- Treat patients equally and with respect
- Provide more information to patients

Finalisation of evaluation hampered by Covid-19

Adapted AwezaMed for **Covid-19**

AwezaMed S2S translation app





Covid-19 app that translates ~1000 English utterances to the other 10 official languages of SA

Supported by 4 multilingual (multimodal) GF applications grammars (no resource grammars)

Which corners to cut?

Workflow for maternal health app (with mRGs)

- Develop abstract and English, Afrikaans concretes
- Create “representative” corpus of English sentences
- Obtain translations in Zulu and Xhosa
- Manually analyse translations with expert linguistic help
- Implement mRGs for Zulu and Xhosa to cover necessary structures
- Implement Zulu and Xhosa concretes (using mRGs)
- Obtain evaluation of randomly generated translations

iterative

Which corners to cut?

Workflow for Covid-10 app (without mRGs)

- Develop abstract and English, Afrikaans concretes

iterative

- Create “representative” corpus of English sentences
- Obtain translations for all other languages
- Analyse Zulu translations with via (m)RG
- Implement Zulu concrete directly
- Bootstrap other languages from Zulu
- Obtain evaluation of randomly generated translations

Which corners to cut?

Zulu workflow with RG

- Develop abstract and English concrete
- Create “representative” corpus of English sentences
- Obtain translations for target language
- Analyse Zulu translations with via RG
- Implement Zulu concrete using RG
- Obtain evaluation of randomly generated translations

iterative

Which corners to cut?

Zulu workflow with RG

- Develop abstract and English concrete
- Create “representative” corpus of English sentences
- Obtain translations for target language
- ***Analyse Zulu translations with via RG***
- ***Implement Zulu concrete using RG***
- Obtain evaluation of randomly generated translations

iterative

Corpus of triples: semantic tree/linearisation/**syntax tree**

NN instead of concrete syntax

Concrete syntax ≈ Mapping between application abstract syntax and RG abstract syntax

Can we *learn* such a mapping?

Can we *create* the data we need?

Try parsing/NLU direction first:

- linearisation/syntax tree → semantic tree
- direction that benefits most from increased robustness

NN instead of concrete syntax

Augmenting data via an application grammar

- Corpus of triples: semantic tree/syntax tree/linearisation
- Augmentation rules:

$$T_A, T_B \rightarrow t_a, t_b$$

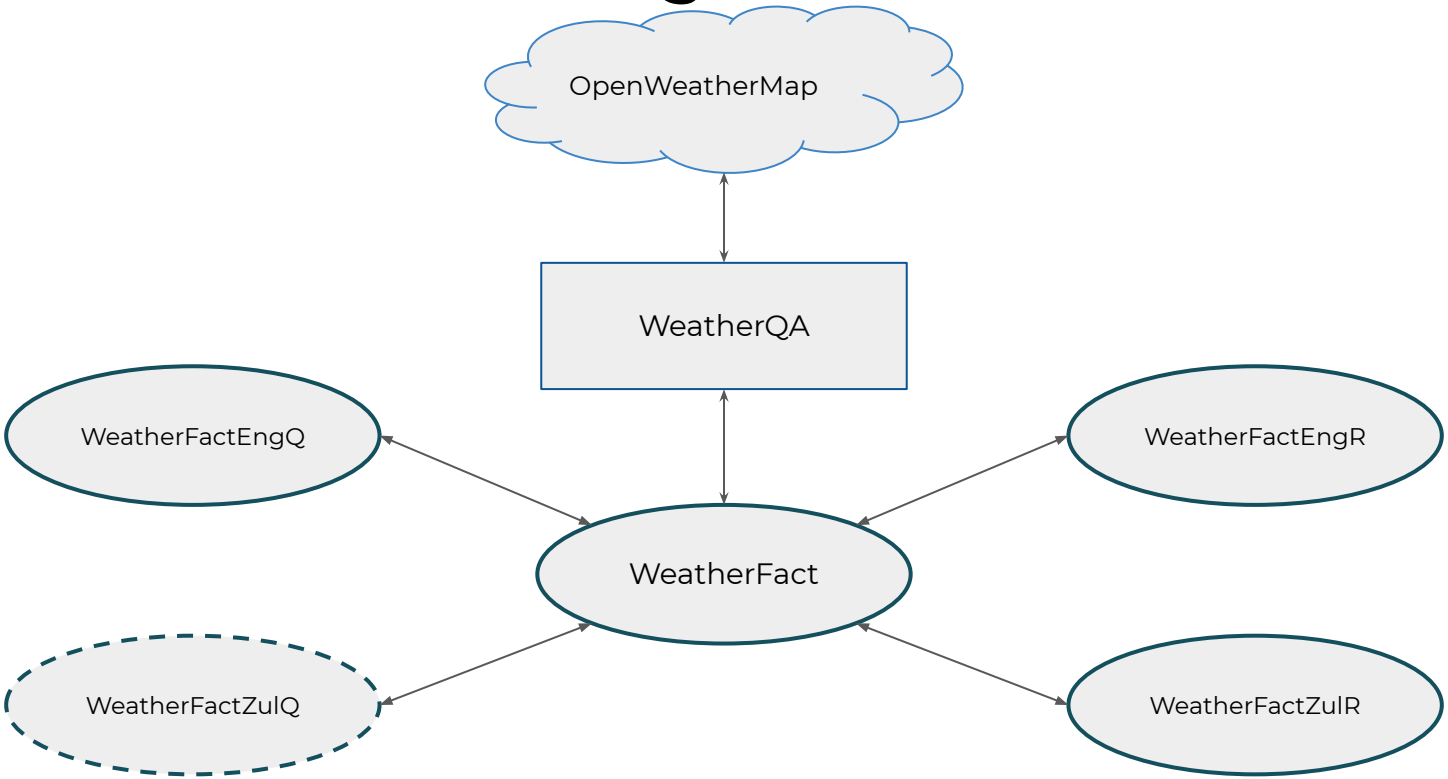
Given a semantic tree T_1 and corresponding syntax tree t_1, if a semantic tree T_2 is acquired by substituting T_A in T_1 for T_B, and a syntax tree t_2 is acquired by substituting t_a in t_1 for t_b, then t_2 is the corresponding syntax tree of T_2.

Question-understanding

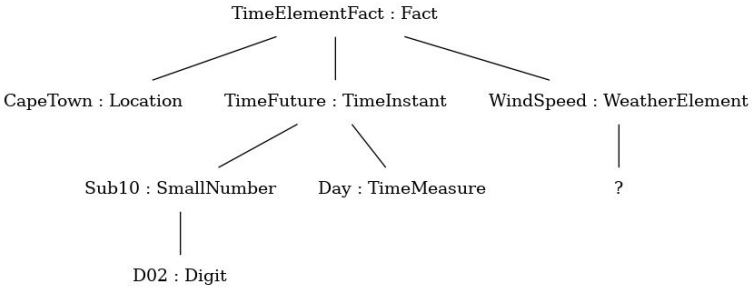
Pivot to a GF-driven question-answering system about weather conditions

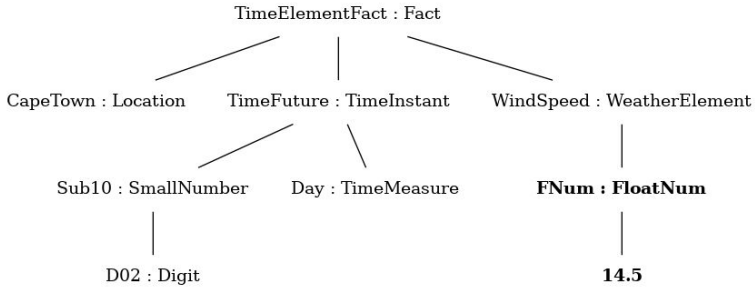
- Somewhat comparable abstract complexity
- Better coverage of numbers (<99), which are linguistically complex in Zulu
- Allows us to focus on NLU

Question-understanding



Question-understanding[1]





Uzobe uvunguza kangakanani umoya ekapa ngemva kwezinsuku ezimbili?

How much will the wind blow in Cape Town in two days?

Isivinini somoya eKapa ngemva kwezinsuku ezimbili sizobe singamakhilomitha ayi-14.5.

The wind speed in Cape Town in two days will be 14.5 km/h.

1. https://github.com/GrammaticalFramework/gf-contrib/tree/master/weather_qa

NN instead of concrete syntax

- Develop abstract and English concrete
- Create representative seed corpus of English sentences
- Obtain two sets of Zulu translations
- Obtain parse trees using Zulu resource grammar
- Augment data to obtain many more semantic-syntax-linearisation triples
- (Evaluate augmentation)
- Linearise augmented syntax trees to obtain semantic-syntax-Zulu triples
- Train NN that maps Zulu/syntax tree to semantic tree
    - try different flavours

NN instead of concrete syntax

Parse translations

- Extend the WIP Zulu RG to obtain 148/152 parses, with at least one for each English sentence

Augment data

- 148 utterance seed corpus → 341 254 utterance augmented corpus

Balance corpus

- Duplicate* some utterances based on size of semantic tree

NN instead of concrete syntax

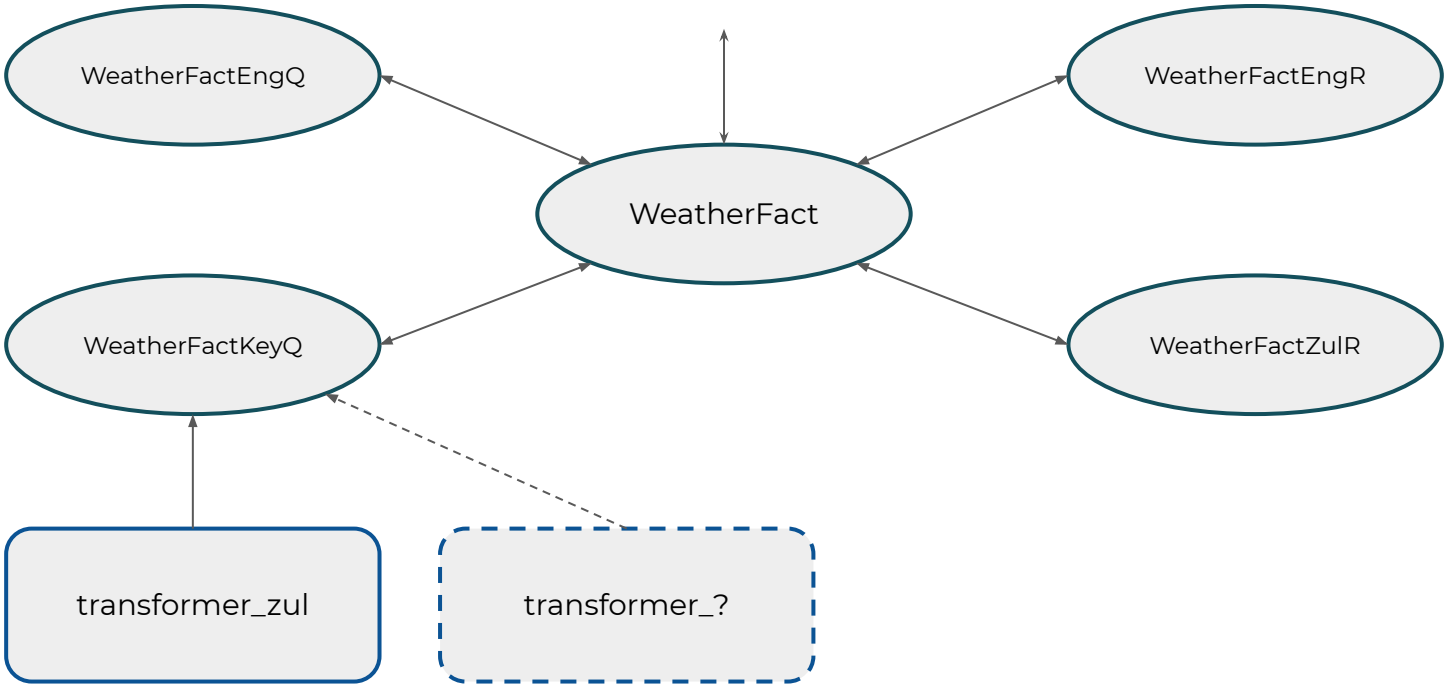
Input flavours

- Zulu linearisation, tokenised on word boundaries
- Zulu linearisation, tokenised on syllable boundaries
- Zulu linearisation, tokenised at character level
- Syntax tree string
- Lemma sequence (lexical functions)

Output flavours

- Semantic tree string eg “TimeElementFact CapeTown (TimeFuture (Sub10 D02) Day) (WindSpeed ?)”
- **Keyword sequence*** eg “CapeTown TimeFuture D02 Day WindSpeed”

NN instead of concrete syntax



NN instead of concrete syntax

Evaluation

- Test set selected from augmented data set
  - Can the NN approximate a concrete syntax?
- New translations
  - Is the NN more robust?

- Preparing an independent test set
  - Seed corpus is minimal and representative ito concepts, but not balanced
  - Augment new translations in the same way (71 parses → 138 896)
  - Randomly select 100 sentences

NN instead of concrete syntax

Model	F score (test)	P score (test)	F score (new)	P score (new)
token2key	1.0	100%	0.65185	33%
syllable2key	1.0	100%	0.83098	**54%**
char2key	0.99537	97.84%	0.81897	52%
syntax2key	1.0	100%	0.81894	52%
lemma2key	0.99966	99.67%	**0.83629**	53%
*lemma2key**	*N/A*	*N/A*	*0.18702*	*0%*

Conclusion

- GF is a good fit for speech and language applications for resource scarce languages
- In South Africa, resource scarcity starts with human resources
- A RG could provide the leverage to generate enough custom data to train models that perform similarly to application grammar concrete syntaxes
    - Less flexible in some ways
    - More flexible in other ways