Overview

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

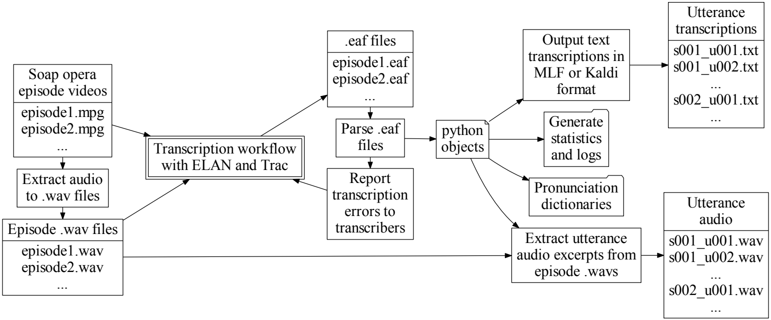
- The South African Corpus of Multilingual Code-switched Soap Opera Speech – Ewald
- Data augmentation by synthesis of code-switched bigrams using word embeddings – Ewald
- Semi-supervised acoustic model training for five-lingual South African code-switched ASR – Astik

The South African Corpus of Multilingual Code-switched Soap Opera Speech

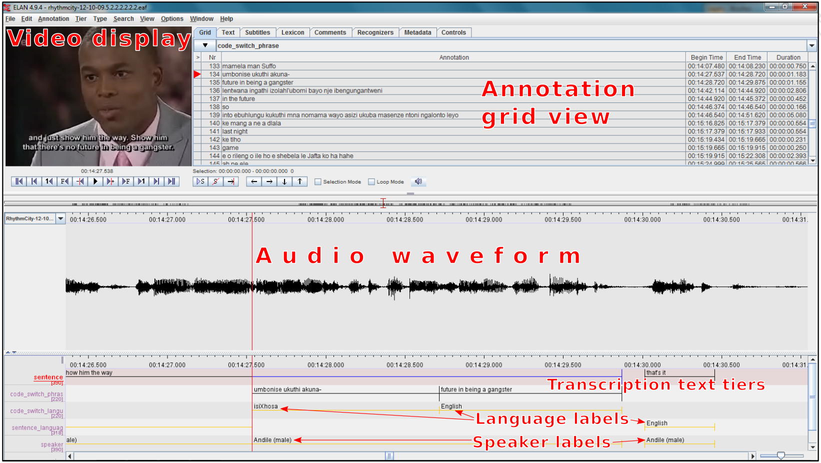
Data overview

- **Data collection**: Time-consuming
- **Data sparsity**: What to do?
  - Collect more data.
  - **Augment existing data.**

Transcription procedure



ELAN media annotation tool



- Transcribed by bilingual speakers
- Five tiers

Annotated example

`sentence:`

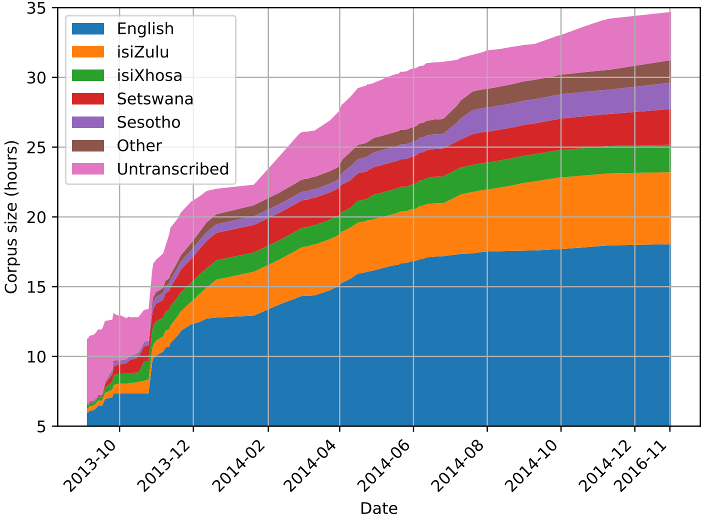
`code_switch_phrase:` *umbonise ukuthi akuna*- future in being a gangster

`code_switch_language:` isiXhosa English

`sentence_language:` <no_segment_or_annotation>

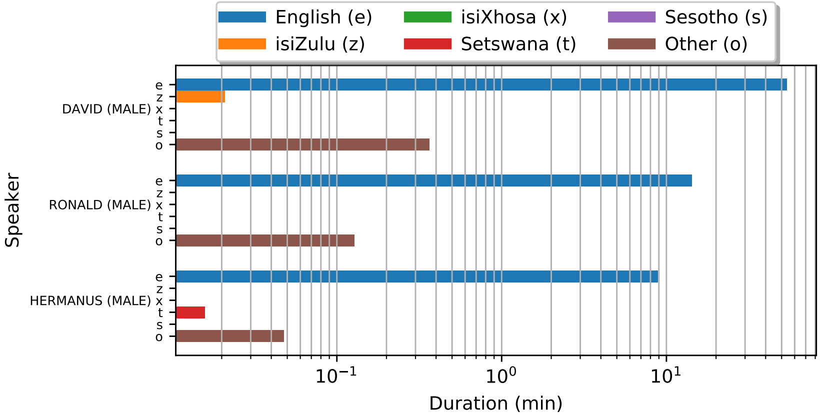
`speaker:` Andile

Corpus growth



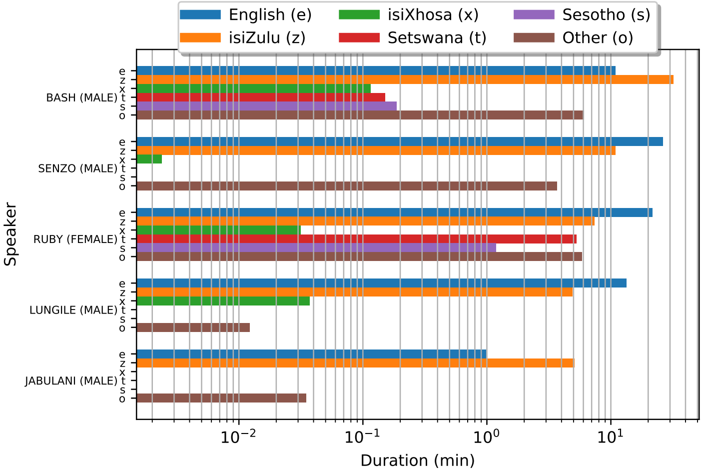
- 35 hours of segmented speech.
- English has the highest occurrence.

Speaker's language distribution – English



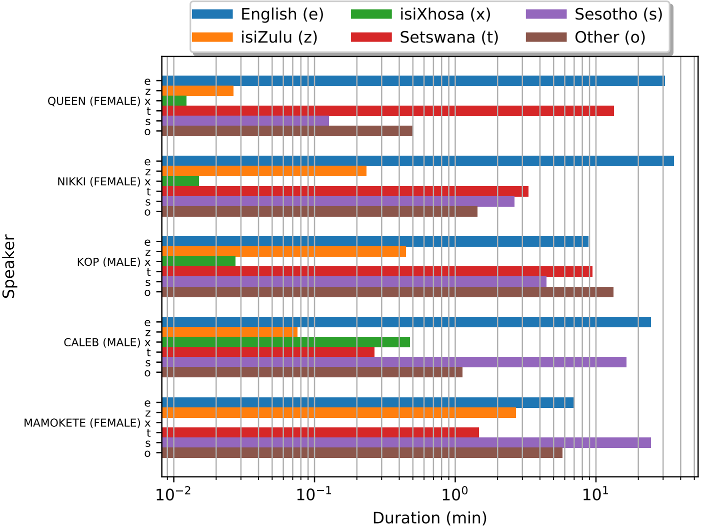
- English speaker hardly code-switch.

Speaker's language distribution – IsiZulu



- Speakers with near even distribution between isiZulu and English.
- Third speaker shows more varied use of language.

Speaker's language distribution – Sotho-Tswana



- Speakers with near even distribution between Setswana, Sesotho and English.

Examples of bigrams with code-switching

English	**IsiZulu**		**IsiZulu**	**English**
understand	-a (verb terminative)		(plural prefix) ama-	shares
English	**IsiXhosa**		**IsiXhosa**	**English**
sister	wakhe (his/her/its)		(plural prefix) izi-	shares
English	**Setswana**		**Setswana**	**English**
know	go (that)		(that) re	you
English	**Sesotho**		**Sesotho**	**English**
feel	-a (verb terminative)		(plural prefix) di-	parents

Corpus analysis

- Code-switched segments are short (250 to 750ms).
- English insertion the most frequent
- 64 to 92% of code-switched bigrams occur only once.
- Code-switched sentences:
  - 50% one switch
  - 1.86 switches per sentence.
- Spontaneous soap opera speech 1.7 times faster than prompted speech.
- English, Nguni and Sotho-Tswana language typologies differ.
  - Agglutinative
  - Conjunctive vs disjunctive orthography

Data augmentation by synthesis of code-switched bigrams using word embeddings

Data augmentation by synthesis of code-switched bigrams using word embeddings

- Use well resourced monolingual English data to synthesis bilingual code-switching examples absent in training data.
- Word embedding
  - Automatically discover semantic, syntactic relationships between words.
  - Words are mapped to a vector space.

Word embedding training

```
Surely those words are not pleasant words to say at any time or in any
situation.
                                   ...
Governments worldwide are also investigating the possibility of implementing
multipurpose citizen cards therefore the S. A. government had no choice but
to issue a new tender conservatively valued at another one billion Rand for a
smart card.
                                   ...
```

	surely	those	words	pleasant	...	investigating
surely	0	58	4	0	...	0
those	58	0	604	1	...	19
words	4	604	0	4	...	0
pleasant	0	1	4	0	...	0
⋮	⋮	⋮	⋮	⋮	⋱	⋮
investigating	0	19	0	0	...	0

surely	−0.315	0.234	0.068	0.019	0.047	0.225	−0.178	...
those	−0.247	0.087	−0.035	−0.076	−0.172	0.124	−0.173	...
words	−0.324	−0.084	0.120	−0.029	−0.035	0.322	0.011	...
pleasant	−0.264	−0.097	−0.141	0.127	−0.325	0.077	0.014	...
investigating	−0.183	0.179	−0.026	−0.134	−0.053	−0.159	−0.019	...

Word embedding querying

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

	isiZulu→English	English→isiZulu
Bigram	i- album	relationship yethu
Trigger	i-	relationship
Target	album	yethu

Word embedding querying

	isiZulu→English		English→isiZulu	
Bigram	i- album		relationship yethu	
Trigger	i-		relationship	
Target	album		yethu	
Query word	album		relationship	
	Similarities	*Cosine score*	*Similarities*	*Cosine score*
Result words	song	0.874	relationships	0.833
	movie	0.806	friendship	0.685
	film	0.774	affair	0.608
	soundtrack	0.771	engagement	0.578
	series	0.736	conversation	0.574
	footage	0.703	environment	0.573
	gig	0.696	image	0.563
	animated	0.689	life	0.556
	record	0.681	lives	0.550
	book	0.672	links	0.536

Word embedding querying

	isiZulu→English	English→isiZulu
Bigram	i- album	relationship yethu
Trigger	i-	relationship
Target	album	yethu
	Synthesised bigrams	
	i- song	relationships yethu
	i- movie	friendship yethu
	i- film	affair yethu
	i- soundtrack	engagement yethu
	i- series	conversation yethu
	i- footage	environment yethu
	i- gig	image yethu
	i- animated	life yethu
	i- record	lives yethu
	i- book	links yethu

Word embedding results

	EZ		EX		ET		ES	
	dev	tst	dev	tst	dev	tst	dev	tst
% unseen before augmentation	83.4	86.9	93.4	95.7	74.5	77.4	81.4	82.7
% unseen after augmentation	71.6	78.4	90.4	92.2	59.4	61.5	70.8	66.6
Factor increase in CSBG types	7.3		7.9		7.8		8.6	
% CSBGs correct: B	22.4	19.3	12.6	12.0	13.1	18.8	15.9	11.8
% CSBGs correct: S	24.4	20.0	14.6	11.4	14.7	19.9	16.0	12.8

Conclusions

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

- Used well-resourced monolingual text to synthesise bilingual code-switching.
- Inclusion of synthesised code-switched bigrams:
  - reduced language model perplexity with up to 31% across a language switch boundary;
  - improved code-switched bigram coverage with up to 21%.
- Improvement in code-switched bigram accuracy in 3 of 4 language pairs.

Semi-supervised acoustic model training for five-lingual South African code-switched ASR

Important ASR evaluation terms

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

- **Perplexity** : Measure of how well a language model predicts the next word, given a sequence of words. **Lower perplexity values = better language models.**
- **Word Error Rate (WER)** : Performance metric for automatic speech recognition (ASR) systems:

$$WER(\%) = \frac{S + D + I}{N} \times 100$$

where N is the total number of words in the reference transcription and S, D and I are substitutions deletions & insertions. **Lower WERs = more accurate recognition.**

Multilingual Corpus for Code-switched South African Speech

Manually segmented and transcribed training speech

Duration in hours (h) and minutes (m) of languages in the unbalanced soap opera corpus. Mono dur: Monolingual duration, CS dur: Code-switched duration

Language	**Mono dur (m)**	**CS dur (m)**	**Total (h)**	**Total (%)**	**Word tokens**	**Lexicon entries**	**Word types**
English	754.96	121.81	14.61	69.26	193 986	8 275	5 965
IsiZulu	92.75	57.41	2.50	11.86	24 387	11 352	7 448
IsiXhosa	65.13	23.83	1.48	7.03	22 313	6 169	5 975
Sesotho	44.65	34.04	1.31	6.22	21 398	2 792	2 437
Setswana	36.92	34.46	1.19	5.64	13 831	1 902	1 625
Total	994.43	271.54	**21.10**	100	275 915	30 489	23 453

Additionally, we have 11 hours of manually segmented but **untrasncribed** soap-opera speech a total of 127 speakers (69 male and 57 female)

Dev and Test sets used to evaluate CS ASR performance

Duration (minutes) of English, isiZulu, isiXhosa, Sesotho, Setswana monolingual (mdur) and code-switched (cdur) utterances

			English-isiZulu (EZ)		
	emdur	zmdur	ecdur	zcdur	**Total**
Dev	0.00	0.00	4.01	3.96	8.00
Test	0.00	0.00	12.76	17.85	30.40
			English-isiXhosa (EX)		
	emdur	xmdur	ecdur	xcdur	**Total**
Dev	2.86	6.48	2.21	2.13	13.68
Test	0.00	0.00	5.56	8.78	14.34
			English-Setswana (ET)		
	emdur	tmdur	ecdur	tcdur	**Total**
Dev	0.76	4.26	4.54	4.27	13.83
Test	0.00	0.00	8.87	8.96	17.83
			English-Sesotho (ES)		
	emdur	smdur	ecdur	scdur	**Total**
Dev	1.09	5.05	3.03	3.59	12.77
Test	0.00	0.00	7.80	7.74	15.54

1 464, 691, 798, & 1 025 language switches observed in the EZ, EX, ES, & ET test sets, **no monolingual test data**

Supervised Training

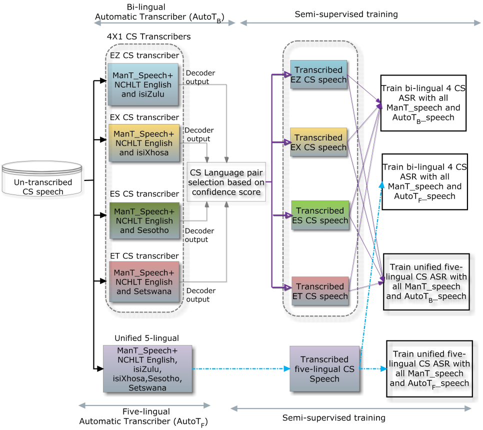
Two main approaches:

1. Bi-lingual CS ASR (can recognize two languages simultaneously)
   - **EZ**: ManT(21.1h) + NCHLT English(>50h) + NCHLT isiZulu(>50h)
   - **EX**: ManT(21.1h) + NCHLT English(>50h) + NCHLT isiXhosa(>50h)
   - **ES**: ManT(21.1h) + NCHLT English(>50h) + NCHLT Sesotho(>50h)
   - **ET**: ManT(21.1h) + NCHLT English(>50h) + NCHLT Setswana(>50h)
2. Five-lingual CS ASR (can recognize five languages simultaneously)
   - **EZXST**: ManT (21.1h) + NCHLT (English + isiZulu +isiXhosa + Sesotho +Setswana) (> 250h)

- **VERY** little soap-opera data available to develop robust CS ASR
- Given the amount of out-of-domain monolingual NCHLT speech, the improvement is not so significant

Semi-supervised Training

Use the best possible Code-Switched ASR to transcribe new soap-opera speech to increase the amount of in-domain acoustic training data.



Semi-supervised training framework for the five-lingual(-·-·-·→) and 4× Bilingual CS (⟶) transcription systems.
(ManT: Manually transcribed; AutoT: Automatically transcribed.)

Automatic Transcriptions

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

23 290 segmented, untranscribed soap opera utterances ($\pm$ 11 h)

- Parallel bilingual code-switch transcription (AutoT$_B$)
  - Each utterance decoded in parallel by each bilingual decoder
  - Output with highest confidence score provides transcription & language pair label
  - $\Rightarrow$ 7 951 EZ, 3 796 EX, 11 415 ES and 128 ET
- Unified five-lingual code-switch transcription (AutoT$_F$)
  - Not restricted to bilingual output
  - Bantu-to-Bantu language code-switching also observed
  - $\Rightarrow$ 3 390 isiZulu, 142 isiXhosa, 657 Setswana, 1 069 Sesotho, 3 952 English & 14 080 CS

Language Modelling: SRILM Toolkit

UNIVERSITEIT STELLENBOSCH UNIVERSITY

- EZ, EX, ES, ET vocabularies contain 11 292, 8 805, 4 233, 4 957 word types, closed with respect to train, development & test sets
- 3-gram LMs: 4 $\times$ bi-lingual, 1 $\times$ 5-lingual

Text resources used for LM development

Type	LM	Text source	
		In-domain	Out-of-domain (Monolingual)
Bi-lingual	EZ	EZ train text	English, isiZulu
	EX	EX train text	English, isiXhosa
	ES	ES train text	English, Sesotho
	ET	ET train text	English, Setswana
5-Lingual	EZXST	EZ, EX, ES, ET train text	English, isiZulu, isiXhosa, Sesotho, Setswana

LM perplexity

MPP: monolingual perplexity
CPP: code-switch perplexity (computed **only** across a language switch)
EB: English to Bantu switch; BE: Bantu to English switch

	Dev	Test	all CPP	all MPP
		Bilingual 3-gram language model		
EZ	425.82	601.69	3 291.95	358.08
EX	352.87	788.81	4 914.45	459.04
ES	151.47	180.47	959.01	121.24
ET	213.34	224.53	70.18	160.40
		Unified five-lingual 3-gram language model		
EZ	599.93	1 007.15	6 708.18	561.80
EX	669.07	1 881.82	15 083.65	1 015.93
ES	365.48	345.35	3 617.44	207.84
ET	236.96	277.48	2 936.63	158.15

- The CPP for EZ and EX are high due to agglutination
- Perplexities of the five-lingual trigram are substantially higher than those of the the bilingual LMs
- isiZulu and isiXhosa are **agglutinative language** $\rightarrow$ high perplexity

Acoustic Modelling: Kaldi Toolkit

- Training sets of all relevant languages combined into a single pool of training data
- Conventional context-dependent GMM-HMM acoustic model used to obtain alignments
- Two different types of neural networks for acoustic modelling:
  1. 11 layers of Factorized Time-Delay Neural Network (TDNN-F)
  2. 2 Convolutional Neural Network (CNN) layers added to 11-layer TDNN-F
- Applied 3-fold data augmentation (1.1x faster, normal, 0.9x slower)
- Features: High resolution MFCCs (40-dimensional, without derivatives), pitch (3-dimensional) & i-vectors for speaker adaptation (100-dimensional)

Acoustic Model Training Pool

AutoT$_B$: transcriptions by **bi-lingual** automatic transcription systems
AutoT$_F$: transcriptions by **five-lingual** automatic transcription system

Type	Target Languages	Training set
Bi-lingual (4xCS)	EZ, EX, ES, ET	ManT (Baseline) ManT + AutoT$_B$ ManT + AutoT$_F$
5-Lingual (1x5CS)	EZXST	ManT (Baseline) ManT + AutoT$_B$ ManT + AutoT$_F$

Bi-lingual CS acoustic models adapted to target language pair after multilingual training

Bi-lingual Semi-Supervised Experiments

Mixed WERs (%) for 4 CS language pairs

CS Pair	Bilingual code-switched ASR							
	TDNN-F (Baseline) ManT		TDNN-F ManT+AutoT$_B$		CNN-TDNN-F ManT+AutoT$_B$		CNN-TDNN-F ManT+AutoT$_F$	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
EZ	41.4	47.5	39.5	44.9	38.2	44.0	36.3	**43.2**
EX	45.7	52.3	42.4	48.7	39.7	47.3	40.4	**46.7**
ES	58.6	60.2	56.3	56.2	54.0	53.6	53.8	**52.9**
ET	54.1	51.0	51.7	50.4	48.5	45.6	47.0	**45.5**
Overall	49.9	**52.7**	47.5	**50.1**	45.1	**47.6**	44.4	**47.1**

- Semi-supervised TDNN-F training using AutoT$_B$ $\rightarrow$ absolute WER reduction of 2.6% relative to baseline
- CNN-TDNN-F $\rightarrow$ additional 2.5% reduction
- Acoustic models retrained with AutoT$_F$ transcriptions $\rightarrow$ best performance

5-lingual Semi-Supervised Experiments

Mixed WERs (%) for 4 CS language pairs

CS Pair	Unified 5-lingual code-switched ASR							
	TDNN-F (Baseline) ManT		TDNN-F ManT+AutoT$_F$		CNN-TDNN-F ManT+AutoT$_F$		CNN-TDNN-F ManT+AutoT$_B$	
	Dev	Test	Dev	Test	Dev	Test	Dev	Test
EZ	39.7	50.3	36.6	46.2	35.8	**44.8**	37.3	47.3
EX	44.5	63.6	43.6	59.9	42.2	60.1	42.3	**59.2**
ES	54.8	50.4	53.5	48.9	53.9	48.8	51.45	**48.2**
ET	48.3	46.0	47.4	43.3	45.1	**42.9**	51.2	49.1
Overall	46.8	**52.6**	45.3	**49.6**	44.2	**49.2**	45.5	**50.9**

- Five-lingual recognition is more difficult since it allows more freedom in terms of permissible language switches
- Semi-supervised TDNN-F training using AutoT$_F$ $\rightarrow$ absolute WER improvement of 3% relative to baseline
- CNN-TDNN-F acoustic model trained with AutoT$_B$ transcription $\rightarrow$ no significant improvement
- Deteriorated performance for EX and EZ due to higher corresponding perplexities values

Summary & Conclusions

UNIVERSITEIT
STELLENBOSCH
UNIVERSITY

- We introduced semi-supervised acoustic model training
- Aim: improve the performance of under-resourced code-switched ASR for four South African language pairs
- 11 hours of manually segmented but untranscribed soap opera speech containing code-switching was processed Bi-lingual & 5-lingual automatic transcription systems
- Results indicate that both approaches were able to reduce overall WER substantially
- 5-lingual system exhibited a bias towards English
- Despite the added confuseability inherent in decoding five languages, the 5-lingual system showed good performance

Thanks for your attention! Any questions?