A First South African Corpus of Multilingual Code-switched Soap Opera Speech

Ewald van der Westhuizen and Thomas Niesler
Stellenbosch University, South Africa

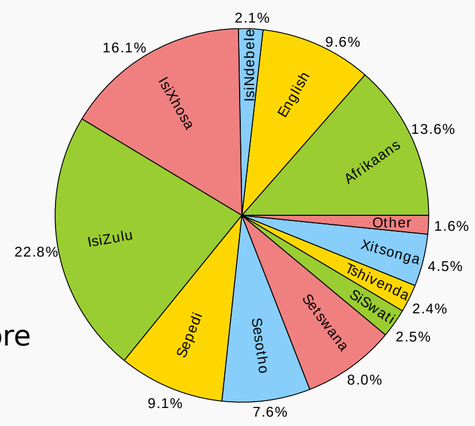
2019 | INTERNATIONAL YEAR OF Indigenous Languages

Summary

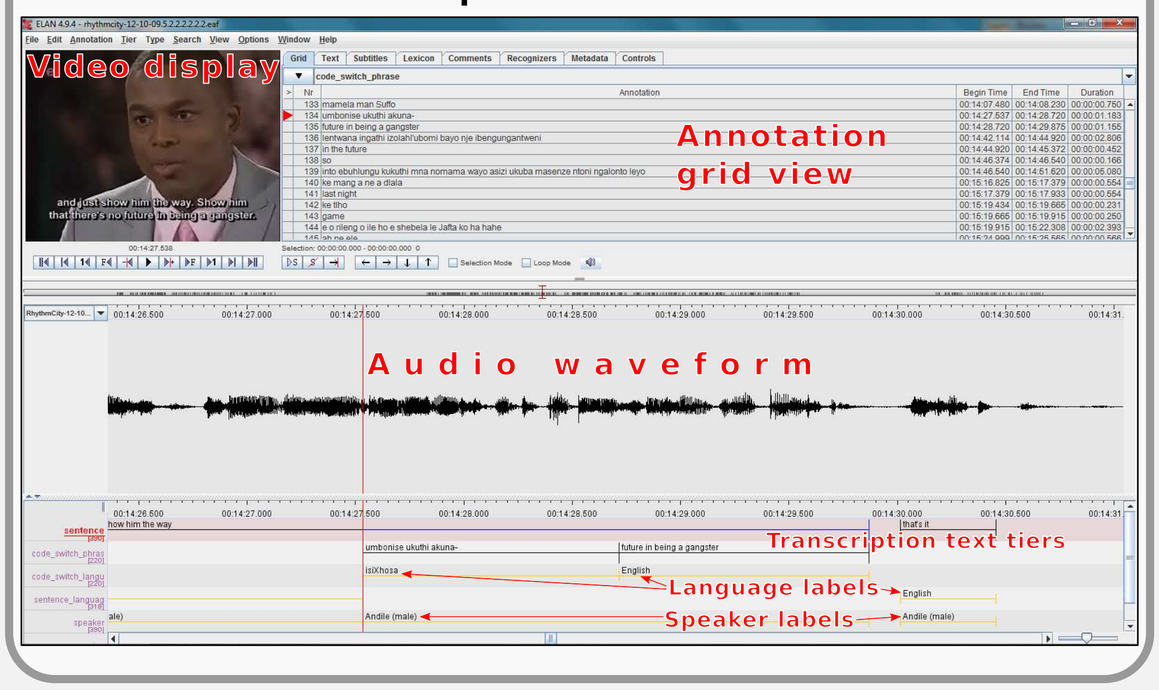
We introduce a speech corpus containing multilingual code-switching compiled from South African soap operas. The corpus contains monolingual as well as code-switched examples of English, isiZulu, isiXhosa, Setswana and Sesotho speech. The last four are indigenous languages, all belonging to the Southern Bantu family. IsiZulu and isiXhosa are Nguni languages that, while distinct, are to some degree mutually intelligible and linguistically similar. The same applies to Setswana and Sesotho, which are Sotho-Tswana languages. The data contains both inter-sentential and intra-sentential code-switching. Intra-sentential code-switching occurs as alternation, insertion as well as intra-word switches.

Background

- South Africa (SA) has 11 official languages and most citizens are multilingual.
- Code-switching occurs when a speaker alternates between languages in a conversation or sentence.
- Code-switching is prevalent in everyday South African speech.
- Speech from SA soap operas display rich examples of code-switching.
- Actors tend to ad-lib. The speech is therefore spontaneous.
- Actors usually switch between English and their mother tongue.



Transcription with ELAN



Statistics and Analysis

- Corpus divided into four language-balanced subcorpora, each with one language pair.

Statistics for: #utt: number of utterances; #tok: number of word tokens; eng: monolingual English duration; neng: monolingual Bantu language duration; ecs: total duration of English segments in code-switched sentences; necs: duration of corresponding Bantu language segments in code-switched sentences.

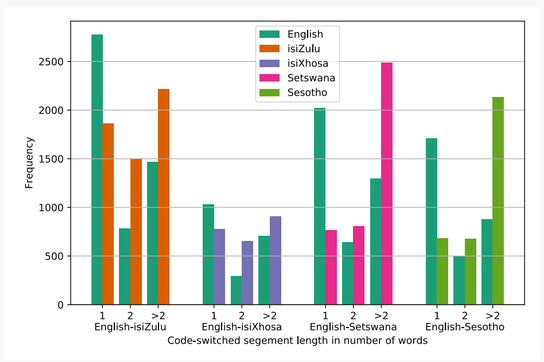
Set	#utt	#tok	eng	neng	ecs	necs	dur
English-isiZulu	9k	59k	1.55h	1.55h	1.04h	1.31h	5.45h
English-isiXhosa	8k	38k	68m	60m	26m	35m	3.14h
English-Setswana	6k	44k	41m	35m	48m	47m	2.86h
English-Sesotho	7k	42k	50m	40m	34m	45m	2.83h

- Language switch directions fairly evenly balanced.
- Intra-word switches from/to Bantu prefixes/suffixes.

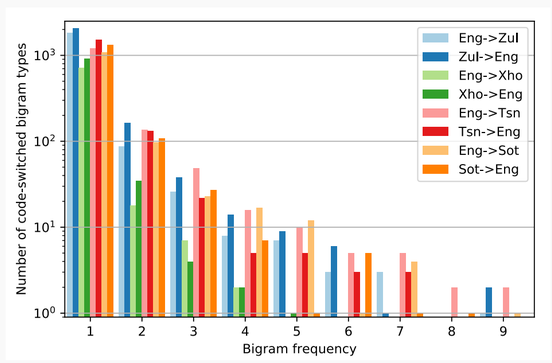
Sw.dir: language switch direction; #sw: number of language switches; #pref/#suff: number of prefixes/suffixes.

Sw.dir	#sw	Sw.dir	#sw	#pref	#suff
Eng⇨Zul	3099	Zul⇨Eng	3717	1930	419
Eng⇨Xho	1195	Xho⇨Eng	1479	566	87
Eng⇨Tsn	2600	Tsn⇨Eng	2728	217	566
Eng⇨Sot	2109	Sot⇨Eng	2280	234	525
Total #sw	19207				

Frequency of monolingual segments consisting of one, two and more than two word tokens found in code-switched sentences for each subcorporus.



Occurrence counts of code-switched bigrams for the various training sets.



- Coverage of bigram types containing code-switching (CSBG) in the training sets with respect to the development and test sets for the subcorpora.

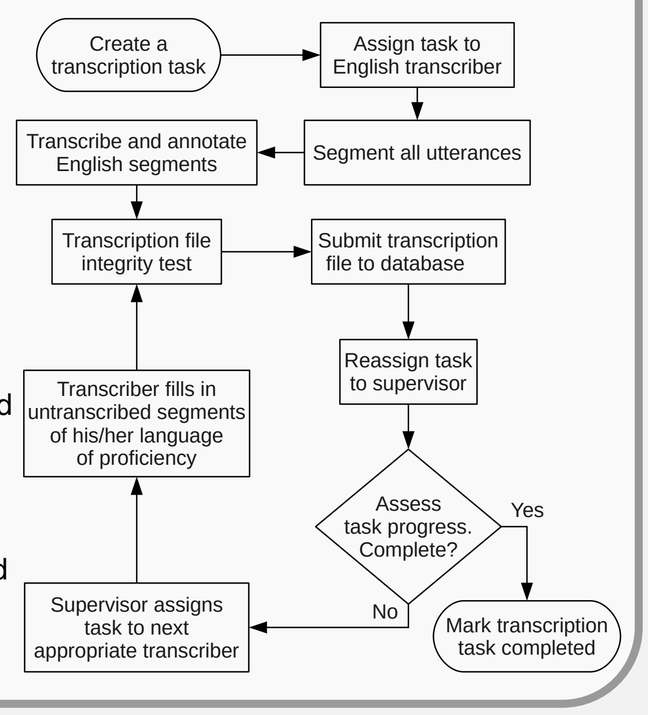
CSBG types: The number of dev/test set bigrams types containing code-switching;
seen CSBG types: The number of dev/test set bigram types containing code-switching which also occur in the training data; % unseen: The percentage of dev/test set code-switched bigram types not occurring in the training data.

	EZ		EX		ET		ES	
	dev	tst	dev	tst	dev	tst	dev	tst
# CSBG types	355	1371	197	668	455	932	312	739
# seen CSBG types	59	180	13	29	116	211	58	128
% unseen	83.4	86.9	93.4	95.7	74.5	77.4	81.4	82.7

- Language topologies differ. English is considered analytic, isiZulu and isiXhosa synthetic and Sotho-Tswana mid-way between analytic-synthetic.
- IsiZulu and isiXhosa use conjunctive orthography, while Sotho-Tswana use disjunctive orthography.
- Insertion of an English word into the Bantu matrix is the most common form of code-switching.
- English insertions may contain Bantu affixes.
- The majority of code-switched bigrams occur only once.
- Postlexical deletion in fast spoken Bantu speech occurs frequently.

Transcription workflow

- Transcription by fluent bilingual speakers.
- Audio collected from 626 soap opera episodes.
- Audio format: 32kHz 16-bit PCM
- Segment audio into sentences.
- Multilingual sentences subdivided into languages.
- English transcribed first.
- Remaining languages transcribed by appropriate bilingual speaker.



Examples from the data

Insertional code-switching
Alternational code-switching
Contracted Bantu words (postlexical deletion)
Bantu language affixes attached to English words

English-isiZulu:
If *ungahambisani nendlela engidla ngayo ngizophindela e-* flat *-ini lami* and take care of myself.
K'khon' ey- sign *-iw' ezansi* and those are paid for.

English-isiXhosa:
And *into endiku-* assure *-isha ngayo kukuthi mna* I'm a phone call away, son.

English-Setswana:
O book *-ile di-* room *tse i-* two.

English-Sesotho:
But I thought miss Stella *ke ena a* arrange *-ang dilo tse ka mokana wa tsona*.

A code-switched bigram consists of a trigger token followed by a target token. For example, in the bigram "*lami* and" from the English-isiZulu example above, "*lami*" is the trigger and "and" the target. The table lists the most and second most frequent triggers and targets in each subcorpus with their frequencies.

Trigger		Target		Trigger		Target	
English ⇨ isiZulu				isiZulu ⇨ English			
and	128	-a	179	i-	465	and	98
no	94	ukuthi	158	u-	205	right	88
English ⇨ isiXhosa				isiXhosa ⇨ English			
and	56	ukuba	34	i-	157	and	46
so	32	yakho	27	u-	44	right	41
English ⇨ Setswana				Setswana ⇨ English			
so	70	-a	234	di	209	you	61
if	62	ke	193	le	178	and	52
English ⇨ Sesotho				Sesotho ⇨ English			
if	50	-a	226	di	175	and	47
so	40	-e	127	ke	128	right	34