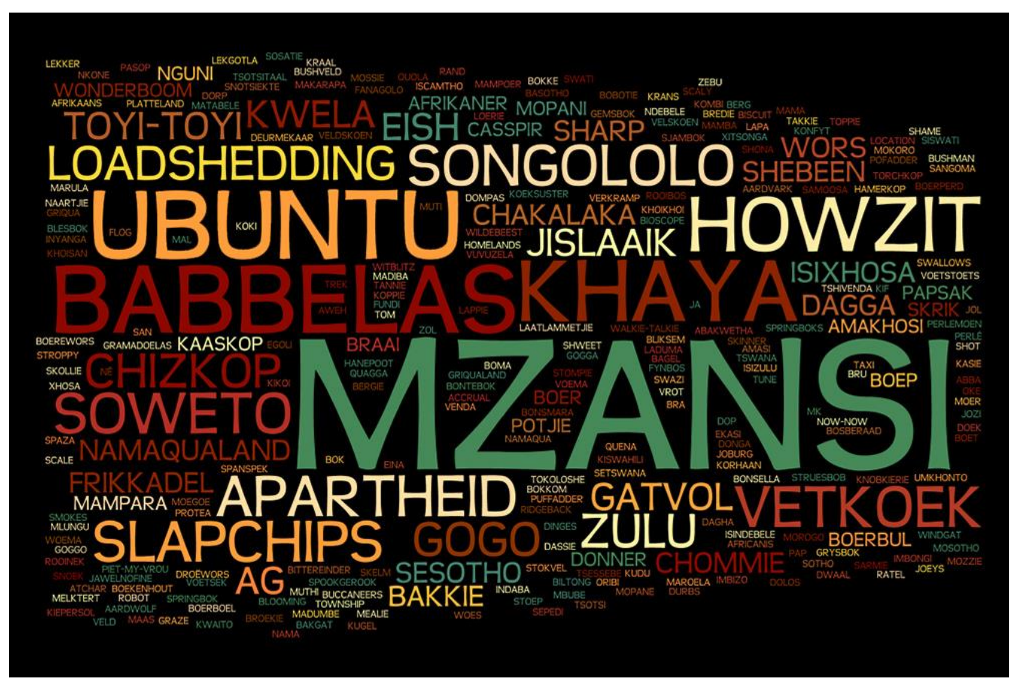
South African English: Creating a Corpus and Part of Speech Tagging Efficiency

Introduction

Part of speech tagging is a mature field with much research against American and British English, but none of the research is specific to South African English. To this end, a corpus of South African English is necessary.



A corpus is a compilation of text that is used for linguistic research. The problem is that South African English doesn't have any electronic corpus and only has smaller spoken (audio) corpora that focus on sub varieties of the language.

Taking these problems into consideration our project aims were:

- Build a comprehensive corpus that fairly represents sub varieties of South African English and can be used for linguistic research.
- Test the efficiency of current POS taggers on South African English to help determine whether a new tagger should be developed specifically for South African English.

Methods and Materials

All data gathered for the corpus was Done by writing Custom web crawlers in Python by using the Scrapy library.

The following data categories were Included as part of our corpus:

- News websites
- Political speeches
- Media statements
- South African Books(Fiction)
- South African Twitter data
- South African Blogs

When it came to testing the efficiency of POS tagging, we tested four different taggers against South African English and compared the results to the original papers on the taggers. The measure of how well the taggers performed was represented by F-score.

Results

	Stanford	HunPos	NLTK	Tree
Daily Voice	0.950	0.956	0.959	0.942
M&G	0.964	0.967	0.967	0.968
Blue Sky	0.95	0.957	0.952	0.960
news24	0.968	0.968	0.97	0.974
Blogs	0.962	0.938	0.934	0.957
Aggregate	0.949	0.955	0.953	0.96

Table 1: Part of speech tagging F-scores

Data Category	Number of texts	Token count
South African twitter data	3287 tweets	50269
South African Blogs	98 articles	59260
Media statements and advisories	101 statements	49850
Political Speeches	106 speeches	182851
South African fiction	10 Books	313887
News websites	803 articles	398966
Total	4405(1118 excluding tweets)	1 055 083

Table 2: Break down of corpus token count

Sub-category	Region	Example where found
National News	National	News24
Local News	KwaZulu-Natal	The Mercury
Sport (News)	Eastern Cape	HeraldLive
Health (News)	National	Mail and guardian
Education (News)	National	Mail and guardian
Lifestyle and Entertainment (News)	Western Cape	Daily Voice
Finance (News)	National	The South African
Travel (Blogs)	Gauteng	Rattle and Mum
Lifestyle (Blogs)	Western Cape	Boring Cape Town Chick
South African fiction (Books)	National	Project Gutenberg

Table 3: Token count for data sub-categories

Discussion

Concerning POS tagging efficiency, the taggers struggled in a few main areas:

- The Stanford Tagger almost always tagged the first word of a sentence as a Proper Noun.
- If a single word was unidentified (marked as foreign), at least 2 more words would automatically be marked as foreign as well.
- Social media jargon was almost never identified correctly
- Possessive markers ('s) was almost always incorrectly identified due to improper tokenization.
- Contractions were incorrectly identified in many cases, also due to incorrect tokenization.

A few key takeaways from the corpus results were:

- Categories such as Blogs and twitter data did not have as much content as other categories due to spelling mistakes and non-English content being part of several tweets and Blogs.
- News websites made up majority of the corpus due to there being widespread availability of sources.

Conclusions

- Comprehensive corpus for SAE was achieved, though not all sub varieties were fairly represented due to the lack of availability of data for some South African provinces.
- The TreeTagger performed the best coming within margin of error, whilst other taggers only performed 1-2 % worse. So in short, we conclude that:

University of Cape Town

Department of Computer Science

Email: dept@cs.uct.ac.za

Tel: 021 650 2663

The Team :

Alec Badenhorst
Email: bdnale004@myuct.ac.za

Muhammed Umar Khan
Email: khnmuh036@myuct.ac.za

Supervisors:

A/Prof. Maria Keet

Email: mkeet@cs.uct.ac.za