Developing a broadband automatic speech recognition system for Afrikaans

FEBE DE WET[1,2], ALTA DE WAAL[1], GERHARD B VAN HUYSSTEEN[3]
[1]Human Language Technology Competency Area, CSIR Meraka Institute, Pretoria, South Africa
[2]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
[3]Centre for Text Technology (CTexT), North-West University, Potchefstroom, South Africa
E-mail: fdwet@csir.co.za

INTRODUCTION

Afrikaans
- Low Franconian, West Germanic language – closely related to Dutch
- One of South Africa's 11 official languages
- Third largest official language (6 million L1, 16 million L2/L3)
- Spoken in South Africa, Namibia & expatriate colonies (Australia, Canada, UK, etc.)
- Under-resourced

This study reports on the development of:
- Broadband Afrikaans speech corpus
- Afrikaans pronunciation dictionary for ASR
- Baseline phone recognition results

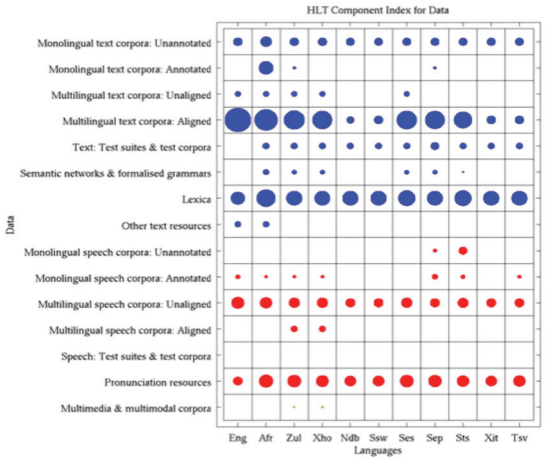

Figure 1: Current HLT status of South Africa's official languages.

HLT RESOURCES FOR AFRIKAANS

Existing resources: Text data
- Currently no corpus big enough to estimate language models reliably
- "Taalkommissiekorpus" (CTexT, 2009)
  - Standard formal Afrikaans in written form
  - Stratified according to stratum of ICE (written)
  - 57 million words, 500,000 unique entries
  - Used to estimate frequency of occurrence for words

New resources: Speech data
- Radio news bulletins
- Archive material (2001-2004) & recordings since end 2010
- 330 bulletins ~= 27 hours of audio data
- Manually transcribed
  - e.g. mispronounced words, unintelligible speech, bulletins, interviews, etc.
  - Speaker gender, language ID, acoustic quality, music/speech ID
- News readers: 18 male, 10 female

New resources: Pronunciation dictionaries
- *Lwazi* dictionary
  - 4,997 entries, 906 *Default & Refine (D&R)* rules
- *Resources for Closely Related Languages* (RCRL) dictionary
  - Bootstrapped with Lwazi
  - Conflicting *D&R* rules were analysed to find systematic errors
  - Expanded incrementally in batches of 5,000 words
  - 24,000 entries, 3,224 Default & Refine *(D&R)* rules
- Background dictionary
  - Words that are conflicting with typical Afrikaans *D&R* rules
  - Proper names and acronyms e.g. Johannesburg, Fifa
  - Abbreviations and initialisms e.g. sms, CNN, CSIR
  - Compounds with initialisms e.g. M-Web, Cell-C
  - Informal words e.g. merc - mercedes, bru - brother
  - Number words e.g. 4x4, media24 and 94.2
  - Initial pronunciation predicted with *D&R* rules, manually verified
  - 6,644 entries including variants

ASR SYSTEM

- MFCC-based system
- 3-state HMM phone models
- Triphone clustering & semitied transforms (final set of triphones)
- Data
  - No speaker overlap between train and test set
  - Used only news reader speech

	Train	Test
Duration (mins)	313.4	48.4
# female speakers	6	4
# male speakers	12	6

Table 1: Size and composition of the training and test sets.

RESULTS

	% Correct	% Accuracy
RCRL APD	75.9	68.7
RCRL APD + background pdict	76.1	69.1

Table 2: Phone-recognition correctness (Correct) and accuracy (Accuracy) with and without a background dictionary.

- Results compare well with previous experiments with telephone speech
- Using background dictionary improves recognition accuracy
- News bulletins contain many proper names – some are in background dictionary, most are not
- *D&R* rules seem to be good at predicting pronunciations for proper names

FUTURE WORK

- Improve pronunciation prediction of proper names
- Use more data without training a speaker-specific ASR system
- Resource, data and technology "recycling" – Afrikaans/Dutch
  - e.g. part-of-speech tagging for Afrikaans bootstrapped with Dutch system
  - e.g. grapheme-and-phoneme-to-phoneme (GP2P) mapping to improve G2P
  - Improved acoustic modelling for Afrikaans using Dutch data

UNIVERSITEIT·STELLENBOSCH·UNIVERSITY
jou kennisvennoot • your knowledge partner

NORTH-WEST UNIVERSITY
YUNIBESITI YA BOKONE-BOPHIRIMA
NOORDWES-UNIVERSITEIT

CSIR
our future through science