The South African Directory Enquiries (SADE) Name Corpus

Jan W.F. Thirion · Charl van Heerden · Oluwapelumi Giwa · Marelie H. Davel

Accepted: 25 January 2019

Abstract We present the design and development of a South African directory enquiries (DE) corpus. It contains audio and orthographic transcriptions of a wide range of South African names produced by first-language speakers of four languages, namely Afrikaans, English, isiZulu and Sesotho. Useful as a resource to understand the effect of name language and speaker language on pronunciation, this is the first corpus to also aim to identify the "intended language": an implicit assumption with regard to word origin made by the speaker of the name. We describe the design, collection, annotation, and verification of the corpus. This includes an analysis of the algorithms used to tag the corpus with meta information that may be beneficial to pronunciation modelling tasks.

Keywords Speech corpus collection · Pronunciation modelling · Speech recognition · Proper names

1 Introduction

Multilingual environments, such as in South Africa, present unique and interesting challenges to systems dealing with pronunciation variability. Spoken dialogue systems need to adequately deal with various factors that affect speech production, such as a speaker's socio-economic background, mother tongue language, age, and gender [37, 2]. Differences among these factors result in speakers producing words with varying pronunciation, leading to deteriorated recognition performance [1].

Hand-crafting rules to deal with pronunciation variation is both time-consuming and impractical. This is particularly the case for resource-scarce environments where the availability of data is limited, time and cost considerations do not always justify the collection of additional data, and skilled linguists may not be readily available [4, 11]. Automatic methods are thus needed to predict pronunciation given the known variables about a speaker. These typically require a corpus of example utterances and their associated phonemic transcriptions. Pronunciation rules can then be trained and used to predict a word's pronunciation given a set of variables such as the speaker's mother tongue language, as well as the language of origin of the word.

Information access by telephone has proven to be convenient and effective to many people, especially those in rural communities. Of particular importance here is the problem of a directory enquiry (DE) system [6, 7, 29]. In a DE application, speakers call an interactive voice response (IVR) system and an automatic speech recogniser (ASR) recognises the name of a company or person uttered by the caller. Information is then relayed to the caller or an automatic call distributor (ACD) system relays the call to the company or person. Such an automated system could result in high cost-savings to a company providing such a service. Dealing with multilingual factors in this environment is crucial, requiring a high-quality task-specific corpus.

In this paper, we present the design, collection and analysis of a South African names corpus to be used to build a DE system, called SADE. We give details around the collection, annotation, and verification of the corpus. In addition, we do pronunciation modelling to derive legitimate variants in a semi-automatic process. Finally, we present the algorithms for the automatic identification of the linguistic origin of words and the

Jan W.F. Thirion
Multilingual Speech Technologies (MuST), Faculty of Engineering, North-West University, South Africa.
E-mail: thirionjwf@gmail.com

Charl van Heerden
E-mail: cvheerden@gmail.com

Oluwapelumi Giwa
E-mail: oluwapelumi.giwa@gmail.com

Marelie H. Davel
E-mail: marelie.davel@nwu.ac.za

concept of intended language tagging. Using these algorithms, a corpus can be tagged with meta information that could be beneficial to pronunciation modelling tasks. Interesting preliminary results are presented.

2 Background

Multilingual speech-enabled systems face the problem of pronunciation variation due to personal and regional factors such as a speaker's mother tongue language, educational level, age and gender [36]. A speaker may also perceive the linguistic origin of a word to be different from its actual language of origin. This results in pronunciation variation and dealing with these factors remains an unsolved problem [21].

A particularly difficult task is that of dealing with the pronunciation variation of proper names, which are the most important and common words for automated call routing, navigation, voice dialling, directory assistance, and auto-attendant applications. This is exceptionally difficult due to proper names typically having different morphological and phonological rules than other words in a language [39], the typically low correlation between orthography and pronunciation [36] and the inability of a speaker to pronounce words in a given language [28]. The number of proper names also exceeds other words significantly [1, 26], which requires innovative algorithms to model pronunciation accurately.

A high-quality names corpus is useful for the development of pronunciation prediction algorithms. As a result, a number of corpora has been developed:

- *ONOMASTICA* [17] - a project to create a multi-language pronunciation dictionary for European names. A total of 11 languages with up to 1 million names per language were included. Letter-to-sound rules and self-learning software formed part of the goals of the project.
- *Autonomata Spoken Names Corpus* [20] - a corpus of 3 540 unique names of Dutch, French, English, Turkish and Moroccan origin. The names include street names, person names (names and surnames) and city names.
- *Multipron* [16] - a corpus of first name and surname combinations where both the names and speakers were selected from four South African languages, namely isiZulu, Sesotho, English and Afrikaans.
- *Multipron-split* [38] - the Multipron corpus reworked to split names and surnames such that items only contain words from a single language of origin, making the development of pronunciation modelling algorithms easier.

Various strategies have been proposed to deal with pronunciation variation:

- Knowledge of the mother tongue of the speaker, as well as the linguistic origin of the word, have been found beneficial to producing better pronunciation variants [25]. The consistency of cross-lingual pronunciation of proper names was studied for four South African languages, namely Afrikaans, English, Setswana and isiZulu [23]. It was confirmed that knowledge of the linguistic origin of the word helped to guide pronunciation thereof. A study on how mother tongue and the linguistic origin of the word affect ASR performance, was performed in [32].
- A combination of grapheme-to-phoneme (G2P) conversion and phoneme-to-phoneme (P2P) conversion [40] have been shown to work well for accent adaptation [27], as well as speech recognition [21, 33, 34]. In the latter, the generation of a canonical pronunciation with a G2P converter was followed by the generation of alternative pronunciations using a P2P converter.

In this paper, we continue to develop the concept of "intended language" introduced in earlier work [23, 24]. A speaker may choose a pronunciation from a set of pronunciations that is dependent not only on the language of origin of the word but also on the language the speaker intends to pronounce the word in. This may or may not be the speaker's perception of the language of origin of the word, and as a result, this form of pronunciation variation could be considered a stylistic variation. As an example, the name "richard" may be pronounced as / r i x a r t / if the speaker intends to produce an Afrikaans pronunciation, or / r\ i t S @ d / if an English pronunciation is intended. We annotate the transcriptions of the audio recordings with the intended languages the speaker may have had in mind when the pronunciation was selected, using an automated process to try and identify the "intended language". We aim to model the intended language explicitly in order to improve the recognition of proper names.

The application area of interest considered here is that of a DE system. It has been found that recognition accuracy decrease logarithmically as more names are added to the vocabulary [22]. A practical DE system needs to recognise more than a million names, and as a result, DE systems have been approached as a large vocabulary continuous speech recognition/recogniser (LVCSR) task [10]. Various strategies have been attempted [35]. Despite remarkable progress, automated DE is not a solved task. Task completion rates of 81% and turn accuracy rates as low as 61% (when multiple recognition attempts are taken into account) have been reported [41].

3 Corpus design

The main goal with the development of the SADE corpus was to collect and analyse audio samples from multilingual speakers producing the type of proper names typically requested in a DE system.

The eleven official languages of South Africa fall into two language families, namely Southern Bantu and Germanic [42]. Two subfamilies, namely Nguni and Sotho-Tswana, subdivide seven of the Southern Bantu languages into two further sets, consisting of languages that display certain similarities. These three groups - Germanic, Nguni and Sotho-Tswana - form the main language families within the list of official languages. Two additional languages are from the Venda and Tswa-Ronga families, respectively. Apart from the official South African languages, many additional languages are spoken and used in South African society, influencing the names of public entities such as restaurants or business names. This creates an interesting mix of both names and possible pronunciations.

The design of the corpus was influenced by the following decisions:

- Using publicly occurring names from a wide variety of sources: Name lists were selected from publicly available information, rather than constructed artificially for the purpose of this corpus.
- Corpus size and speaker contribution: a goal was set to collect speech from 40 individuals, each contributing 400 utterances per speaker. This was based on an analysis in [3] where it was found that (a) when building speech recognition systems, phonetic content is more important than speaker diversity, and (b) that good recognition accuracy is available within these parameters. (While speech recognition accuracy was not our goal here, our corpus analysis and annotation process required speech recognition capabilities.)
- Balancing corpus size and speaker diversity: Analysis was restricted to four speaker languages to be able to obtain statistically useful information from the proposed corpus size.
- Four diverse languages were selected in order to obtain a wide range of speaker effects. The four languages selected - Sesotho (a Sotho-Tswana language), isiZulu (a Nguni language), English and Afrikaans - are the four most commonly spoken in Gauteng, and represent the three main language families.
- Data annotation: in order to investigate the interplay between word language, speaker language, and the concept of an "intended" language, all speaker utterances were tagged with the applicable languages at a word level. In order to do this in as an efficient way as possible, automated suggestions were reviewed and corrected manually by language practitioners.

More detail concerning prompt construction, respondent selection and the data collection and annotation protocol is provided in Section 4, below.

4 Data collection

This section provides detail with regard to the collection of the data, the selection of prompts and respondents, quality control performed, as well as the collection platform and protocol. We also highlight some of the practical issues we found during the data collection process.

4.1 Prompt selection

Prompts were selected from a combination of various Internet queries, as well as personal names from the North-West University academic register and other names collected by volunteers. The names were mixed randomly so that no single web source would be compromised. Web sources included the following:

- Well known book titles and authors;
- Famous South Africans;
- Popular songs (artists / titles);
- Fortune 500 companies;
- Google trends queries;
- Popular movies;
- South African municipality names; and
- South African restaurants names and location.

From this collection of proper names, we constructed 80 unique prompt sheets (2 sessions for each of 40 speakers). Each prompt sheet contained 27% unique prompts, and 63% prompts which could be overlapping among speakers. While only 400 utterances were required per speaker, additional utterances were collected to be able to discard problematic utterances where needed. To be specific: each sheet contained 220 prompts (names or name phrases). Of the 440 prompts recorded per speaker, 120 prompts were uniquely recorded by a single speaker, resulting in 4 800 unique prompts. The remaining 320 prompts per session were randomly selected from a larger set of approximately 5 000 prompts.

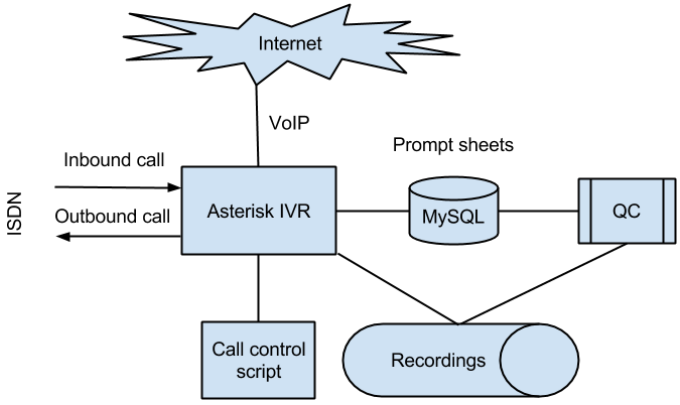

Fig. 1: Overview of the collection platform.

4.2 Collection platform

Our collection platform consisted of an Asterisk interactive voice response (IVR) system with an ISDN telephony card. Recordings were collected over a telephone channel, recorded in 8kHz sampling rate, 16-bit linear PCM format (transcoded from A-law). The IVR was controlled via a programmable script and allows for dynamic user interaction. The caller's progress was also recorded allowing a caller to stop at any time and continue at a later stage. Fig. 1 shows the system architecture:

Callers were allowed to call a number, the IVR call control script instructed the IVR to answer the call and proceed with the interaction with the caller. In the collection process, we realised that telephony costs are still relatively high. To move away from refunding respondents for their telephone call costs, we implemented a callback system. Callers would phone a number and the IVR would then drop the call (similar to a missed call). The IVR then proceeded to call the respondent back on the number they called from and the interaction continued just as if the caller had called in to the system. In addition, we also allowed voice-over-IP (VoIP) calls into the IVR (using the IAX2 and SIP protocols). This allowed us to better monitor and control costs.

We made provision for doing off-line quality control, utilising resources from the NCHLT project [5]. We generated initial phonetic transcriptions for the prompts using a dictionary based on the NCHLT English dictionary. The Quality Control (QC) process ran separately from the IVR and scanned the database for recordings not yet processed. For every recording to be processed, an HTK speech recogniser with a phone loop grammar was used to obtain an initial phonetic transcription of the recording. We used the NCHLT acoustic models, although they were not channel-matched to the environment. Every recording transcription was then aligned with the prompt transcription using Phone-based Dynamic Programming (PDP) [12]. The PDP score and recognition result were both stored for further analysis. We also added the option of doing on-line QC, while there was interaction with the caller. After every nth recording, a stored procedure would calculate the average score of recordings made and decide if the session with the caller should continue or not. If only a few recordings were deemed as having bad alignment scores, these were re-recorded at the end of the session. As a result of the channel mismatch on the acoustic models, we did not enable this feature for the full SADE collection process, but would like to investigate this further in future work.

4.3 Respondent selection

We set a goal of finding at least ten speakers per language with an equal split of males and females (totalling 20 male and 20 female speakers). Only adults in the age range of 18 to 60 were selected. Despite the fact that South Africa has a wide diversity of speakers, we focused on speakers from Gauteng, North-West and Western Province to obtain representative samples of the four focal languages. Volunteer speakers were recruited from within the university; additional respondents were paid for every completed session.

4.4 Collection protocol

The collection protocol consisted of the following process:

- The speakers were e-mailed instructions, which included a phone number, prompt sheet and their speaker identification (ID) number.
- Speakers were required to sign a consent form (as included in Appendix B).
- Every prompt sheet had session information to be completed by the speaker:

  - Prompt sheet number (top right hand corner).
  - Speaker age (e.g. 19).
  - Speaker gender (male / female).
  - Telephony channel (landline, cellphone, VoIP).
  - Provider (e.g. Telkom, Vodacom, Cell C, 8ta, MTN).
  - Handset (e.g. iPhone, Galaxy S3, Nokia Lumia).
  - Date.
  - Time started.
- Speakers were required to be in a quiet environment and have the prompt sheet in front of them.
- Speakers were required to read through the prompt sheet once to familiarise themselves with the prompts.
- When ready to start recording, a number was called and once it rang, the caller was required to drop the call. The system then called the speaker back. Alternatively, if a speaker decided to call the inbound number directly, the recording session would commence. This process allowed for free data collection from a caller perspective. Callers could call the system via landline, mobile phone or VoIP.
- The rest of the recording process was managed by telephony application. It prompted a caller for the prompt sheet number. The caller read the prompts from the sheet after the system played the prompt number and a beep. The system automatically timed out on recordings if no audio was received and automatically moved on to the next prompt. Callers could request to redo a prompt or skip a prompt.
- Instructions were played at regular intervals to the caller and progress indicated to them.
- The caller could drop the call at any time and continue at a later stage using just their speaker ID and prompt sheet ID.
- All prompt sheets had a unique session ID, which consisted of the prompt sheet number, speaker number, and a modulo-10 checksum. This ensured that invalid session IDs entered could be detected. Only one prompt sheet could be used in a session, and prompt sheets were not shared between speakers.

We found that the data collection protocol we followed resulted in most of the logged data matching the prompt sheets as intended, facilitating automated error detection. This was in part due to the modulo-10 checksum used to prevent errors during the data entry phase by the caller.

5 Data annotation

In this section we describe the data annotation and quality verification process in more detail. Note that the entire annotation process (Sections 5.2 to 5.6) was completed first, before evaluating the quality of the resulting corpus (Section 5.7).

5.1 Annotation tags

The corpus was tagged with meta-information that was expected to be useful during pronunciation modelling. These tags are defined in Table 1. Annotations were made at the word level, since many phrases (such as restaurant names) combine words from multiple languages in one phrase.

Tag	**Definition**	**Example**	
Prompt	Orthographic form of prompt exactly as displayed to the user.	Caledon Hotel	Caledon Hotel
Word	Specific word that was being tagged, as it occurred in the prompt.	Caledon	Caledon
Pronunciation	Phonemic pronunciation of the word using SAMPA. (See Table 13).	/ k { l @ d @ n/	/ k a l i d O n /
Speaker language	First language of the respondent.	English	English
Word language(s)	Language(s) of origin of the word. Specifically, the original language community in which the word was used, as well as languages into which it has been incorporated fully. A word may have multiple languages of origin.	English, Afrikaans	English, Afrikaans
Intended language	Also referred to as the "pronunciation language", this is the specific language for which the pronunciation conventions most closely match the pronunciation produced.	English	Afrikaans

Table 1: Tags used to annotate the corpus.

The "word language" is a surprisingly difficult tag to define unambiguously. In most multilingual societies, code-switching (whereby words from other languages are embedded in the primary or "matrix" language) is a frequent phenomenon [30]. Some words then become incorporated in the primary language over time, often also changing their spelling or pronunciation to better fit the conventions of the primary language. This linguistic incorporation process - from pure code-switching to loaned words to full integration - is a gradual one, with no strict boundaries between events. For the purpose of this corpus, when tagging word language, we did not consider frequently occurring code-switched words as being part of the primary language; however, we did include words that are fully integrated into the language. For example:

- The English digit "seven" used in an isiZulu sentence would be tagged as English, not isiZulu.
- The word "zulu", which has become a standard word in English, would be tagged as both isiZulu and English.
- The word "zoeloe" which is the Afrikaans spelling of the word "zulu", would only be tagged as "Afrikaans", not as English or isiZulu also.

Of these, the middle category was most controversial, with individual linguists differing in opinion. For this corpus, we tended towards including more rather than fewer languages where uncertain.

5.2 Overview of annotation process

Of the tags described in Section 5.1 above, only the speaker language could be determined in a straightforward manner, based directly on the information supplied by each respondent. For the rest of the tags, tagging was an iterative process: in the first round a best guess of all parameters (transcript, intended language, word language, pronunciation) was used to initialise the process. Results were analysed using ASR-based techniques and low-confidence annotations reviewed; these were improved through manual verification and the process repeated.

Each of the individual annotation steps is described in more detail below. These are:

- Utterance verification (Section 5.3). The goal of this step was to identify whether the collected audio matched the prompt provided to the respondent.
- Word language identification (Section 5.4): This step aimed to associate each word with its most probable language of origin.
- Pronunciation annotation (Section 5.5): Here the phonemic transcription of words were both generated and verified.
- Intended language identification (Section 5.6): This step aimed to identify the most likely "intended language" of each sample, based on the annotated pronunciation, speaker language, and word language.

The phone set (used both for initial utterance verification and final tagging) was based on the Lwazi phone set [5]. In a process similar to the one used for the *Multipron* corpus [18], the original phone set was further simplified to make cross-lingual transcription possible. As discussed in [18], this was necessary as differences that are phonemic in one language and not in another are often produced fairly arbitrarily by speakers of those languages for which the difference is not phonemic. (Typical examples include duration or aspiration.) The phone set used for all pronunciation annotations is included in Appendix A.

5.3 Utterance verification

Before any further analysis, each utterance was verified individually. The goal was to determine whether the prompting text provided to the respondent matched the captured audio. If so, the prompt was assumed to be the transcription; if not, the utterance was simply discarded.

The audio / prompt match was verified using Phone-based Dynamic Programming (PDP) scores [12]. PDP scores were generated from the output of a speech recogniser: audio was (a) decoded using a flat phone-loop grammar (a grammar that assumed that any phone can follow any phone with equal probability) and (b) the same audio was aligned to the expected pronunciations on a phone-by-phone basis. By comparing the outputs of these two processes, problematic utterances could be identified and discarded.

In general, PDP scores can be calculated using a standard (or "flat") scoring matrix or one optimised for the specific data set. A flat scoring matrix assumes that all phone substitutions have the same cost; a trained scoring matrix assumes that some phone substitutions (some due to speech recognition errors) are more likely than others and should carry less weight when identifying possible errors. In Figures 2 and 3, results obtained using a flat and trained scoring matrix are compared. All utterances are arranged on the x axis according to PDP score, highest (best) scores first.

Scores obtained with the trained matrix are more accurate when identifying actual errors [12], but scores from the flat matrix are more intuitive to interpret. See for example, Figure 2, where the first 10,000 utterances

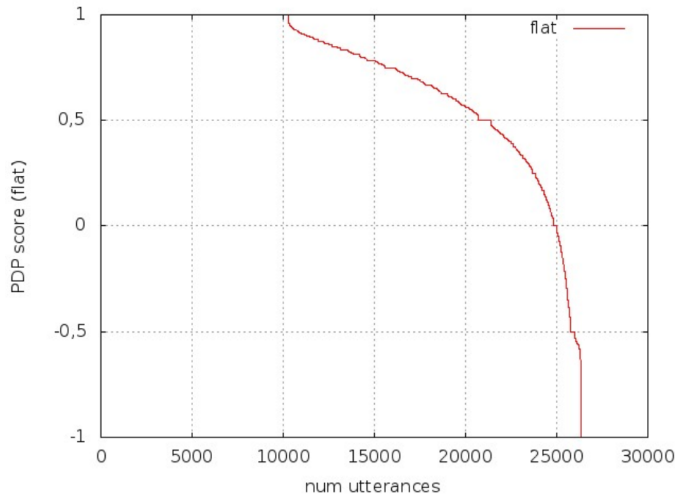

Fig. 2: PDP score per utterance obtained with a flat scoring matrix ordered from best to worst.

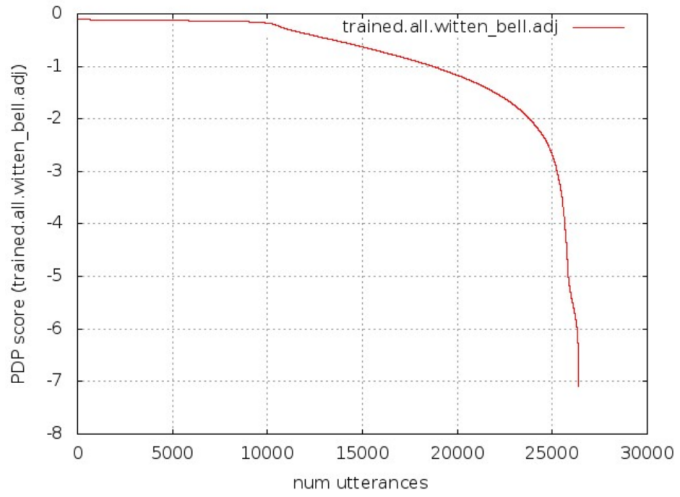

Fig. 3: PDP score per utterance obtained with a flat scoring matrix, ordered from best to worst.

obtained a score of 1: for these utterances, the decoded and aligned phone strings matched exactly. The remainder of utterances shows an increasing mismatch, which could have been caused by various factors: inaccurate initial pronunciations, audio artefacts, or speaker errors. Very low PDP scores (the tail to the right of both figures) are associated with problematic utterances. Typically, utterances prior to the inflection point are usable.

We used a trained scoring matrix to make our selections and did not accept any utterances with a PDP score below -1.5. (See Figure 3). This analysis was repeated on a per-speaker basis, initially flagging the top 420 scoring utterances per speaker for further tagging and analysis. Once all tagging was completed, only the top 400 utterances per speaker were retained in the final corpus.

5.4 Word language identification

For this task we aimed to associate each word with its most probable language of origin. Identifying the language of origin of names has been found to be important when predicting the possible pronunciation of names [25, 9].

Automated text-based language identification (T-LID) of names is a difficult task because names can be short and may have idiosyncratic spellings. There is also often ambiguity in the origin of names, as one name can stem from different language origins. Often, names incorporated in the primary language change their spelling or pronunciation to fit the conventions of the primary language better. This linguistic incorporation process involves a gradual one, in which no strict boundaries exist between events. In this section we refer to the source language of a name as the most likely language from which a name originated - this means that the name was first used and typically follows the spelling systems of that specific language. In predicting language tags, we did not consider frequently occurring code-switched words as being part of the primary language; however, words that are fully integrated into the language were included.

In our work, we used three different techniques to produce preliminary language tags for all words. The results from the different tagging processes were then compared, and where there was disagreement, words were flagged for manual verification, as discussed in more detail below.

5.4.1 T-LID using existing word lists

The first T-LID technique was based on existing lists of known words in South African languages. The list of words was extracted from the original prompts and evaluated against the words in existing word lists (the NCHLT text corpora, as well as public word lists obtained from the Internet - assumed to be less accurate) and existing dictionaries (specifically, the Lwazi and NCHLT dictionaries). As the dictionary-based word lists tended to be more accurate, words were first evaluated against the more accurate word lists. Only if the word was not found in these, it would then be further evaluated against the less accurate word lists.

5.4.2 T-LID using JSMs

The second approach used Joint Sequence Models (JSMs) [8] to perform the T-LID task. JSMs were developed in the context of pronunciation modelling (grapheme-to-phoneme prediction). The algorithm generated graphones (a sequence of graphemes linked to a phoneme sequence) that provided different co-segmentations between a word and its pronunciation. The graphone inventory was optimised using training data. During prediction, the trained models were used to evaluate all possible co-segmentations; the single pronunciation with the highest probability was selected. (The joint probability for each segmentation was estimated by summing over the graphone sequence.) We applied JSMs to the T-LID task, using the process described in [14]: we recast the T-LID task as a G2P task, and applied the standard JSM training process. This experiment added word boundary markers to the training and testing data. Also, we employed log-probability voting to select the final language of origin of a word as described in [14]. Training data was obtained from the NCHLT-inlang [13, 5] dictionaries, providing 15,000 unique words per language. Data was preprocessed by removing words with fewer than two characters, converting all characters to lowercase and removing any loan words that are easily identifiable based on any foreign characters included in the word. In a final round of preprocessing, any multilingual words (the same word included in more than one language list) were removed. Table 3 lists the number of words per language in the training corpus after preprocessing and multilingual word removal.

From this data set shown in Table 2, a balanced corpus was selected as training data shown in Table 3, while randomly extracting 150 words per language as development set and folded back after training. At 1,650 words, this was sufficient for discount estimation, based on the analysis in [8]. Note that specified training set sizes include the development set and indicate the number of training samples *per language* as well.

Language	Word counts	Average word length
Afrikaans	14 413	11.81
English	13 662	7.61
isiNdebele	10 249	11.74
Sepedi	14 285	9.25
Sesotho	9 896	9.87
Siswati	13 289	11.79
Setswana	10 472	10.21
Tshivenda	12 916	9.33
isiXhosa	12 814	10.76
isiZulu	8 690	11.15
Xitsonga	13 028	9.63

Table 2: Word counts per language in the NCHLT corpus after preprocessing and multilingual word removal.

This implementation of JSMs produced a single T-LID prediction for each word. While a variant of the JSM technique is available that produces multiple language candidates [15], that was not applied as part of the current analysis.

5.4.3 T-LID using web info

In our last approach, we experimented with two different techniques using web information, referred to here as "Google Detect" and "Google Translate". Both techniques employ the Google translation systems to identify language identity for those languages that have sufficient web presence. Google Detect[1] uses the built-in functionality provided by the Google Translate API[2] to guess the LID of any input string provided by the user.

[1] https://cloud.google.com/translate/docs/reference/rest

[2] Google Translate API was developed by Google as a proprietary application based on statistical machine translation.

Language	Word counts	Average word length
Afrikaans	8 690	11.59
English	8 690	7.78
isiNdebele	8 690	11.40
Sepedi	8 690	9.35
Sesotho	8 690	9.80
Siswati	8 690	11.80
Setswana	8 690	10.51
Tshivenda	8 690	9.12
isiXhosa	8 690	10.76
isiZulu	8 690	11.00
Xitsonga	8 690	9.52

Table 3: Training data extracted from the NCHLT dictionaries, after processing.

For our task at hand, we used Google Detect to produce a single T-LID prediction for each word, although the system has the capability to produce multiple T-LID predictions.

The "Google Translate" technique to T-LID first guesses the LID of a word and then uses one or more of a set of "proxy languages" to determine whether the word is translatable. Proxy languages are typically world languages such as English, French or Spanish; any number can be selected. If the number of times the word is changed when translated to a proxy language exceeds a threshold, then this is considered a good indication that the word is indeed in the original language. While a low count of translated words does not confirm that the word does not exist in the original language, success provides a strong indication that it does.

In summary, these two web-based techniques, although benefiting from enough data, are only applicable to languages that have sufficient web presence.

5.4.4 Manual review and editing

The steps listed below summarise the ways in which final SADE word lists with their corresponding language tags were generated. During the course of the project, language tags were manually reviewed and corrected based on the following reasons:

- Initially, all words not found in existing word lists were classified using the JSMs and forwarded for a preliminary manual review. This was performed by a single language practitioner who was able to identify obvious mistakes but who was not an expert in all languages occurring in the corpus.
- In the next stage, results from the above methods (word list, JSMs and web-information) were compared: any words that had not been previously verified, or for which at least two of the proposed methods disagreed in terms of T-LID tag, were flagged for further manual confirmation or correction. The word list was, therefore, partitioned into a "Phase 1 tagged" and a "Phase 2 tagged" list. The Phase 1 list consisted of words where at least two of the above techniques agreed on language tags while the "Phase 2" list contained otherwise. The word list belonging to Phase 1 list was incorporated into the corpus while the Phase 2 list was sent for manual review to various practitioners, where language experts represented all 11 South African languages.

The final corpus was tagged using the combination of Phase 1 and Phase 2 results. As the T-LID tag is both ambiguous and the process semi-automatic, we expect to release updates to the language tags over time.

5.4.5 Manual validation

Once the corpus was completed, 600 words were sent for manual validation to determine the accuracy of the entire process. Three hundred (300) words were selected from each of the Phase 1 and Phase 2 tagged lists and forwarded to volunteers for review. Again, the focus was on identifying the accuracy of tags containing South African languages. Standard precision, recall, and the combined F-measure were used to evaluate performance per partition. Precision is the ratio of relevant language tags to the retrieved tags, while recall, on the other hand, is the ratio of relevant language tags retrieved to the total counts of relevant tags.

Good accuracies were observed, as shown in Table 4. However, note that, of the 600 words analysed, results are calculated for 582 words only, as verifiers were not able to tag 18 of the test words. These words contained idiosyncratic spellings (for example, the word "kream"), making LID difficult, even for human verifiers.

5.5 Pronunciation annotation

The next task was to determine and verify the phonemic transcript (or pronunciation) of each utterance. A pronunciation was generated for each word token individually. (One word could have multiple pronunciations.) The process, described in more detail below, consists of the following steps:

Tagged list	words in test set	Precision	Recall	F-measure
Phase 1	300 (291)	99.69	95.54	97.57
Phase 2	300 (291)	99.50	99.50	99.50

Table 4: Final LID accuracy estimate based on manual validation.

1. Initial pronunciations are generated using existing resources, specifically existing pronunciation dictionaries and grapheme-to-phoneme (G2P) rules extracted from these dictionaries.
2. During a first manual verification, the standard pronunciation of words (based on a written version of the words) are verified. Not all words are verified at this stage: audio data is used to analyse and flag possible problematic utterances and to suggest additional pronunciation variants. These pronunciations are reviewed by a linguist, and corrected where necessary.
3. At this stage, words are only associated with their canonical pronunciations. Automated pronunciation variant modelling is now used to generate additional, possibly cross-lingual, pronunciations.
4. The new set of possible pronunciations is again evaluated against the available audio, and a new set of possible problematic pronunciations flagged. This time audio-supported manual verification is performed: the verifier listens to the specific sample and annotates the pronunciation of that specific word.

   Steps (3) and (4) can be repeated until further errors flagged during audio analysis are considered acceptable.

5.5.1 Initial pronunciations

Initial pronunciations were generated using publicly available pronunciation dictionaries (Lwazi, NCHLT) and Default & Refine rules extracted from these dictionaries. For each word, an Afrikaans, English, Sesotho and isiZulu pronunciation were generated (irrespective of the actual word language.) These general pronunciations were remapped to the SADE phone set (see Appendix A) to generate a first pronunciation lexicon.

5.5.2 First manual verification: canonical pronunciations

The same PDP process (described in Section 5.3) was used to flag potentially problematic pronunciations. The standard PDP process does not only allow word samples to be scored but also automatically suggests a possible pronunciation variant based on the audio sample. For each word, we then counted (1) the number of times the word was observed, (2) the number of times each dictionary entry was observed, and (3) the number of times additional variants were suggested. Note that while audio was used to flag these errors, the manual verifier was asked to concentrate on providing canonical pronunciations, rather than possible cross-lingual variants.

As all analysis was ASR-based, the purpose of this part of the process was to improve the accuracy of the ASR system to the extent that more detailed analysis would be possible.

5.5.3 Pronunciation variant modelling

The goal of this step was to introduce additional legitimate variants using a semi-automated process. This was a non-trivial task as it is difficult to differentiate automatically between (a) phonological changes typically modelled at the acoustic level, (b) recognition errors and (c) actual speaker variation. As an example consider the two words in Table 5.

Word	Canonical pronunciation	Observed pronunciations
twenty	t w E n t i	t w E n t i
		t w E n i
simon	s a i m @ n	s a i m @ n
		s i m O n
		s a i m O n

Table 5: Examples of pronunciation variants that are modelled at different levels during ASR system development.

The word “twenty” is sometimes produced with the /t/ observable, and sometimes elision takes place, and the /t/ is not produced at all. This is an example of a normal phonological change that occurs during continuous speech. We did not want to capture this effect in the pronunciation dictionary: instead we captured such variation in the acoustic models (which learnt from the training data that in the /n-t-i/ context, the /t/ is sometimes very short or missing).

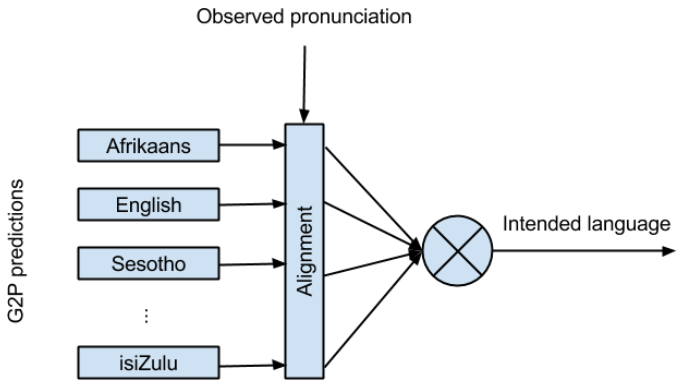

Fig. 4: Overview of the intended language identification algorithm.

On the other hand, the word "simon" is observed as having three distinct pronunciations, which on closer analysis seems to be influenced by the language of the speaker. We *do* want to model these effects. Differentiating between case (a) such as "twenty" and (b) such as "simon" is further complicated by phone recognition errors. While less problematic for words that occur often, it is more difficult to identify the difference between a pronunciation variant and a recognition error for rarer words.

We addressed this issue by focussing on phone substitutions. We used the possible pronunciation variants generated by PDP, and aligned these to the canonical pronunciations. By counting the number of P2P substitutions observed, we created a candidate list of consistent substitutions. These were then applied to the entire dictionary, and the PDP alignment-and-scoring process repeated. This process was quick and semi-manual, and repeated a number of times, before selecting the tail of the distribution for manual review. Note that at this point, the phonemic transcription was reviewed against the word orthography, not yet against its audio form. The overall process at this point (score, select, review, update the acoustic models, re-score) was repeated twice.

5.5.4 Audio-supported manual verification

In the final steps, actual audio samples were manually reviewed, and remaining problematic pronunciations corrected. At this stage, it was expected that problematic audio samples (that is, samples that contained speaker errors or inaudible text) had already been discarded. If not, they were discarded here: pronunciations were only verified/corrected for usable audio.

5.6 Intended language identification

Intended language identification refers to the process of determining the language that a speaker (most probably) had in mind when producing a pronunciation for a word. During this process, we adopted the following hypotheses:

- The intended language can be determined from the speaker's pronunciation of a word;
- If the pronunciation of a word is compared to generally accepted pronunciations, the intended language will be the language of the pronunciation that matches the observed pronunciation best; and
- A language-specific grapheme-to-phoneme (G2P) converter's prediction of a word is a typical pronunciation of the word in that language.

An overview of the intended language tagging process can be seen in Figure 4. Grapheme-to-phoneme (G2P) converters were used to predict typical or reference pronunciations for each of the languages we are interested in. These were then aligned with the observed pronunciation that the speaker produced. Next, the alignment scores were compared to determine the intended language as the language of the G2P converter for which the prediction best matched the observed pronunciation.

In our implementation, we continued to use dynamic programming to align the observed and reference pronunciations. Log probability scores were calculated during the alignment step and as a result, we could determine the intended language using:

$$x = \arg\max_{j} \psi(o|r(x_j)), j = 1 \ldots N_L$$

where x is then considered to be the intended language of an observation, o is the observed pronunciation (string of phonemes produced by the speaker), $r(x_j)$ is the G2P (reference) pronunciation if language x_j is

assumed, and N_L is the total number of languages considered as possible intended languages. The log likelihood of an observation, $\psi(o|r(x_j))$, is determined by calculating:

$$\psi(o|r(x_j)) = \sum_{i=1}^{n_a} \log P(o_i|r_i(x_j))$$

where n_a is the maximum length of the alignment between o and r and $\log P(o_i|r_i(x_j))$ are log likelihood probabilities estimated directly from the data. In this work we used estimates obtained from the confidence matrices trained with the PDP algorithm [12]. We trained a matrix across all languages but (in related work) are also experimenting with language-pair matrices, and retraining the G2P converters based on our assumed intended languages.

5.7 Use in a directory enquiries system

The SADE corpus was used to develop the multilingual directory enquiries application presented in [18]. The Kaldi decoder [31] was used to generate lattices, which were used to obtain confidence scores for hypotheses by doing minimum Bayes risk (MBR) decoding. Acoustic models were trained on a combination of the Lwazi corpora [3, 19].

Almost all speakers achieved a high term recognition accuracy if the confidence was high. The error rates are summarised in Table 6 and Table 7 for the municipality and yes/no grammars respectively.

Confidence	Number of utterances	Term error rate
High	327	3.67
Medium	43	3.26
Low	48	62.50

Table 6: Term error rate for the high, medium and low confidence hypotheses using the municipality recognition grammar.

Confidence	Number of utterances	Term error rate
High	23	26.09
Low	11	54.54

Table 7: Term error rate for the high and low confidence hypotheses using the Yes/No recognition grammar.

The primary question posed during usability testing was: "Is a user able to obtain the number of a specific municipality, using only the SADE system?" The tasks completion rates for both the directed and undirected tasks are provided in Table 8 for the two phases of the usability tests (group 1 and 2). The system achieved an overall task completion rate of 95.65% with the large majority of users finding the system simple and easy to use.

	Directed tasks	Open tasks
Group 1	93.33	90.00
Group 2	95.88	98.82

Table 8: Average task completion rate.

6 Corpus description

What does the language content of the corpus look like? We analyse the different types of words found in the corpus, using the tags from SADE version 1.2.

If words observed in the corpus are combined according to the categories of Table 9, the number of unique words per language category is shown in Fig. 5. If a word can belong to more than one language category, it

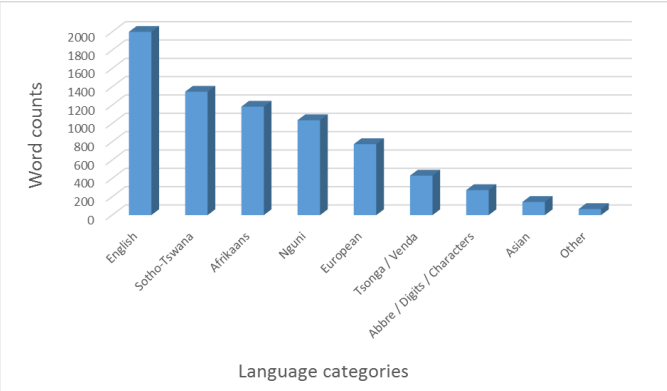

Fig. 5: Number of unique words per language grouping in the v1.2 SADE Corpus.

is counted in each of the applicable categories. This is why the total number of unique word and language tag pairs in Fig.5 is more than the number of unique tokens in Table 10. Three special categories - abbreviations, digits and single characters - were individually marked. As expected, English words dominate the corpus. Note that the y-axis is only shown up to 2 000, even though the number of English words exceeds this number (with a word count of 6 300).

Nguni	Siswati, isiNdebele, isiXhosa and isiZulu
Sotho-Tswana	Sepedi, Sesotho, Setswana
Tsonga/Venda	Xitsonga, Tshivenda
European	Italian, Spanish, Portuguese, Finnish, Turkish, Russian, French, Greek, Dutch, German, Swedish, Latin
Asian	Asian, Chinese, Japanese, Indian, Hindu, Korean and Sanskrit
Other	Unknown, Swahili, Khoisan, Hebrew, Arabic

Table 9: List of languages per language category.

Table 10 summarises the size of the corpus with respect to additional variables: speakers, utterances, prompts and words. The number of mono-, bi- and multilingual words are estimated based on unique words observed: the total number of monolingual words is calculated by counting the number of unique words with only one associated language; “Bi” and “multi” represent unique words with either only two, or and three and more languages associated, respectively. Many of the prompts are multi-word phrases (for example, “Cape Town Hotel School Restaurant”) with some of the words re-occurring in other prompts.

Number of unique monolingual words	6 912
Number of unique bilingual words	1 409
Number of unique multilingual (3+ languages) words	552
Total number of unique words	8 873
Average number of pronunciation variants per word	1.1
Number of unique prompts	9 033
Number of speakers	40
Number of recorded utterances	16 000

Table 10: Basic corpus statistics.

Table 11 shows the relationship between intended language and other language parameters, such as speaker language and word language. This table illustrates to what extent intended language correlates with other language parameters in the corpus. For this analysis, estimates are calculated based on number of tokens. That is, a single word pronounced by different speakers can result in different speaker language and intended language tags. In Table 11, we observe a high correlation between intended language and word language, and a low correlation between intended language and speaker language. The high frequency with which intended language and word language agree (21 678) can be attributed to the fact that a high number of the respondents were familiar with most words in the given utterances. Finally, Table 12 shows intended language distribution according to word occurrence.

The corpus is structured in directories according to speaker language:

	Number word tokens
WL equal to SL equal to IL	6 471
WL equal to IL; IL not equal to SL	21 678
WL not equal to SL; SL equal to IL	1 701
WL not equal to SL not equal to IL	4 056

Table 11: Number of times the word language (WL), speaker language (SL) and intended language (IL) either equals one another or are different.

	Number of unique words	Examples
Words only occurring once	4 306	Zuze, yoga, xate, wow, volume, umjindi
Words with single IL	8 277	&, 007, 1, 1 000, 10 000
Words with 2 different IL	551	Aaron, abo, addams, addis, addo
Words with 3+ different IL	45	Abdullah, absa, bistro, bosh, bourne

Table 12: Intended language (IL) distribution.

```
sade
| README.
| -- data
     | -- <speaker language>
          | -- <speaker>
               | -- audio
               | -- transcriptions
| sade.tagged.<version_no>.txt
```

The meta information related to the entries is captured in a tags file on a *per word* basis. Each entry contains the utterance identifier, orthographic transcription of the word, broad phonemic transcription, list of possibly valid word language(s), the primary speaker language, intended language and the speaker identifier.

File names were chosen to contain some meta information and also allow all files to share a common subdirectory, according to the following format:

```
sade_<lang>_<gender><spk_id>_<num>.wav
sade_<lang>_<gender><spk_id>_<num>.txt
<lang> = zul | afr | sot | eng
<gender> = M | F
```

For example:

```
sade
| -- data
     | -- zul
          | -- sade_zul_F031
               | -- audio
                    | -- sade_zul_F031_001.wav
                    | -- sade_zul_F031_002.wav
                    | -- sade_zul_F031_003.wav
                      ...
               | -- transcriptions
                    | -- sade_zul_F031_001.txt
                    | -- sade_zul_F031_002.txt
                    | -- sade_zul_F031_003.txt
                      ...
```

7 Conclusion

The SADE corpus is a new resource developed in support of multilingual pronunciation modelling research, and the first corpus specifically developed in the application domain of South African directory enquiries. It contains transcribed audio of typical directory enquiry queries (in a variety of languages), generated by 40 speakers (20 male, 20 female) of 4 South African languages. In addition, each word as produced by each speaker is tagged with additional information useful for research on multilingual pronunciations of proper names and general multilingual research.

Corpus annotation resulted in more than 30 000 individually tagged audio samples, with each audio sample associated with a transcription, word language, speaker language and intended language. For three of these tags (pronunciation, word language and intended language) novel algorithms were developed to annotate the corpus in a consistent and efficient manner. Possible errors were flagged for human verification, and the final corpus evaluated by human verification of a set of randomly selected samples. The corpus has been made available freely under an Open Content licence[3]. It is our hope that this resource will stimulate future research into this particularly challenging domain.

Acknowledgements

This work is based on research supported by the Department of Arts and Culture (DAC) of the government of South Africa, through their Human Language Technologies (HLT) unit, and the National Research Foundation (NRF). Any opinion, finding and conclusion or recommendation expressed in this material is that of the authors and the NRF does not accept any liability in this regard. The support by both institutions is gratefully acknowledged.

Appendices

A **The SADE phone set**

Table 13 provides a description of the SADE phoneme set as used to annotate the SADE corpus. For each phoneme, the corresponding IPA and X-SAMPA symbols are also provided.

Description	IPA	X-SAMPA	SADE
Affricates			
Voiceless alveo-labial affricate	tɸ	tp\	t f
Ejective alveolar lateral affricate	tɬˈ	tK_>	t K
Voiceless aspirated alveolar lateral affricate	tɬʰ	tK_h	t K
Voiced alveolar lateral affricate	dɬ	dK	d K
Voiced aspirated alveolar lateral affricate	ɮ	dK\	d K\
Post-alveolar affricate	t ʃ	tS	t S
Ejective post-alveolar affricate	tʃˈ	tS_>	t S
Aspirated post-alveolar affricate	tʃʰ	tS_h	t S
Ejective alveolar affricate	tsˈ	ts_>	t s
Aspirated alveolar affricate	tsʰ	ts_h	t s
Ejective velar affricate	kxˈ	kx_>	k x
Voiceless velar affricate	kx	kx	k x
Voiceless velo-alveolar lateral affricate	kɬ	kK_>	k K
Voiceless labio-alveolar ejective affricate	psˈ	ps_>	p s
Aspirated labio-alveolar affricate	psʰ	ps_h	p s
Aspirated labio-palatal affricate	pʃʰ	pS_h	p S
Voiceless labio-palatal ejective affricate	pʃˈ	pS_>	p S
Voiced post-alveolar affricate	ʤ	d_0Z	d Z
Voiced alveolar affricate	dz	dz	d z
Voiced alveo-labial affricate	dβ	dB	d v
Voiced retroflex affricate	dz̻	dz‘	d z w
Voiced aspirated retroflex affricate	dz̻ʱ	dz‘h\	d z w
Voiced aspirated palatal affricate	ʤh	dZh\	d Z
Fricatives			
Voiceless labiodental fricative	f	f	f
Voiced labiodental fricative	v	v	v
Voiceless dental fricative	θ	T	T
Voiced dental fricative	ð	D	D
Voiceless alveolar fricative	s	s	s

Continued on next column

[3] ISLRN 510-842-952-534-8, available from http://hdl.handle.net/20.500.12185/378 under a Creative Commons Attribution License (3.0 Unported).

Continued from previous column

Description	IPA	X-SAMPA	SADE
Voiced alveolar fricative	z	z	z
Voiceless post-alveolar fricative	ʃ	S	S
Voiced post-alveolar fricative	ʒ	Z	Z
Voiceless velar fricative	x	x	x
Voiceless glottal fricative	h	h	h
Voiced glottal fricative	ɦ	h\	h
Voiceless alveolar lateral fricative	ɬ	K	K
Voiced alveolar lateral fricative	ɮ	K\	K\
Voiced velar fricative	ɣ	G	g x
Voiced bilabial fricative	β	B	v
Voiceless bilabial fricative	ɸ	p\	p
Voiceless retroflex fricative	ʂ	s‘	s w
Voiceless labialised alveolar fricative	sw	sw	s w
Voiced retroflex fricative	zw	zw	z w
Bilabial alveolar fricative	ɸs	p\s	f s
Voiceless bilabial postalveolar fricative	ɸS	p\S	f S
Voiced bilabial postalveolar fricative	βʒ	BZ	v Z
Labiodantal-postalveolar fricative	fʃ	fS	f S
Clicks			
Dental click	ǀ	|	|
Voiced dental click	ǀg$_i$	|\g_0	| g
Aspirated dental click	ǀ^h	|\h	|
Alveolar lateral click	ǁ	||\	||\
Voiced alveolar lateral click	ǁg$_i$	||\g_0	||\ g
Aspirated alveolar lateral click	ǁ^h	||\h	||\
Palatal click	ǃ	!\	!\
Voiced palatal click	ǃg$_i$	!\g_0	!\ g
Aspirated palatal click	ǃ^h	!\h	!\
Stops			
Voiceless bilabial plosive	p	p	p
Aspirated bilabial plosive	p^h	p_h	p
Ejective bilabial plosive	p$'$	p_>	p
Voiceless palatalised bilabial ejective	pj$'$	pj_>	p j
Voiceless aspirated palatalised bilabial plosive	pj^h	pj_h	p j
Voiced bilabial plosive	b	b	b
Voiced bilabial implosive	ɓ	b_<	b
Voiced palatalised bilabial plosive	bj	bj	b j
Voiceless alveolar plosive	t	t	t
Aspirated alveolar plosive	t^h	t_h	t
Ejective alveolar plosive	t$'$	t_>	t
Voiceless palatalised alveolar plosive	tj	tj	t j
Voiceless aspirated palatalised alveolar plosive	tj^h	tj_h	t j
Voiced alveolar plosive	d	d	d
Voiced palatalised alveolar plosive	dj	dj	d j
Voiced aspirated alveolar plosive	dɦ	dh\	d
Ejective palatal plosive	c$'$	c_>	t j
Aspirated palatal plosive	c^h	c_h	t j
Voiced palatal plosive	ɟ	J\	n j
Voiceless velar plosive	k	k	k
Voiceless aspirated velar plosive	k^h	k_h	k
Ejective velar plosive	k$'$	k_>	k
Voiced velar plosive	g	g	g
Voiced aspirated velar plosive	gɦ	gh\	g

Continued on next column

Continued from previous column

Description	IPA	X-SAMPA	SADE
Voiceless alveolar lateral ejective	tlˈ	tl_>	t l
Voiceless aspirated alveolar lateral plosive	tlʰ	tl_h	t l
Nasals			
Bilabial nasal	m	m	m
Alveolar nasal	n	n	n
Palatal nasal	ɲ	J	n j
Velar nasal	N	N	N
Retroflex nasal	ɳ	n'	n w
Voiced aspirated bilabial nasal	mɦ	m_h	m
Voiced aspirated alveolar nasal	nɦ	n_h	n
Labiodental nasal	ɱ	m_j	m j
Voiced aspirated palatal nasal	ɲɦ	J_h	n j
Vowels			
Unrounded high front vowel	i	i	i
Unrounded high front vowel with duration	iː	i:	i
Rounded high front vowel	y	y	y
Rounded high back vowel	u	u	u
Rounded high back vowel with duration	uː	u:	u
Unrounded near-high near-front vowel	ɪ	I	i
Rounded mid-high front vowel with duration	øː	2:	i 9
Unrounded mid-low front vowel	ɛ	E	E
Rounded mid-low central vowel with duration	ɜː	3:	9
Rounded mid-low back vowel	ɔ	O	O
Rounded mid-low back vowel with duration	ɔː	O:	O
Unrounded low front vowel	a	a	a
Unrounded low back vowel with duration	ɑː	A:	A:
Central vowel (schwa)	ə	@	@
Unrounded mid-low front vowel	æ	{	{
Rounded near-high near-back vowel	ʊ	U	u
Rounded mid-low front vowel	œ	9	9
Rounded low back vowel	ɒ	Q	Q
Diphthongs			
	əi	@i	@ i
	œy	9y	9 i
	ai	ai	a i
	ɔi	Oi	O i
	əu	@u	@ u
	au	au	a u
	iə	i@	i @
	uə	u@	u @
	eə	e@	E @
Trills and Flaps			
Alveolar trill	r	r	r
Aspirated alveolar trill	rɦ	rh\	r
Approximants			
Alveolar lateral approximant	l	l	l
Alveolar approximant	ɹ	r\	r
Palatal approximant	j	j	j
Voiced labio-velar approximant	w	w	w
Retroflex lateral approximant	ɭ	l‘	l w

Concluded

Table 13: SADE phone set.

B **Consent form**

The following consent form was used during data collection:

SADE CORPUS COLLECTION CONSENT FORM

I, _____ hereby agree to provide my services in respect of speech data recordings, by reading a list of 660 words or phrases (220 each during three sessions) in a plain and normal tone.

1) I understand that these recordings will be used for the development of various language technologies, including next generation Web applications, and will become an open source resource for the National Centre for Human Language Technologies.
2) I understand that some phrases in the list of words may be offensive and that I have the right not to read these words.
3) I understand that the list of words and phrases is not in any way representative of the interviewer, the project manager, InSyst Labs staff or any of the latterÊijs agents.
4) I agree that I have no right on potential proceeds generated by the recordings.
5) I understand that the recordings will be treated as anonymous data.
6) I understand that I will be paid a once-off total amount of R___ upon successfully completing all three sessions.
7) I understand that "successfully" as defined in (6) is based on my ability to correctly read the prompts.
8) I understand that the 3 prompt sheets I receive are confidential and may not be shared with anyone else.

Signature : _____ Date : _____

References

1. Adda-Decker M, Lamel L (2006) Multilingual Dictionaries. In: Schultz T, Kirchoff K (eds) Multilingual Speech Processing, Academic Press, Berlington, MA, USA, chap 5, pp 123–166
2. Amdal I, Fosler-Lussier E (2003) Pronunciation Variation Modeling in Automatic Speech Recognition. Telektronikk, Telenor pp 70–82
3. Barnard E, Davel M, van Heerden C (2009) ASR corpus design for resource-scarce languages. In: Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), Brighton, UK, pp 2847–2850
4. Barnard E, Davel MH, van Huyssteen GB (2010) Speech technology for information access: a South African case study. In: Proc. AAAI Spring Symposium on Artificial Intelligence for Development (AAI-D), pp 8–13
5. Barnard E, Davel MH, van Heerden CJ, De Wet F, Badenhorst J (2014) The NCHLT speech corpus of the South African languages. In: Proc. of the 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), St. Peterburg, Russia, pp 194–200
6. Bechet F, De Mori R, Subsol G (2001) Very large vocabulary proper name recognition for directory assistance. In: IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp 222–225
7. Bechet F, De Mori R, Subsol G (2002) Dynamic generation of proper name pronunciations for directory assistance. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol 1, pp I–745–I–748
8. Bisani M, Ney H (2008) Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50(5):434–451, DOI 10.1016/j.specom.2008.01.002
9. Church KW (1985) Stress assignment in letter-to-sound rules for speech synthesis. The Journal of the Acoustical Society of America 78(S1):S7–S7
10. Córdoba R, San-Segundo R, Montero JM, Colás J, Ferreiros J, Macías-Guarasa J, Pardo JM (2001) An interactive directory assistance service for Spanish with large-vocabulary recognition. In: Proc. 2nd Annual Conference of the International Speech Communication Association (INTERSPEECH), Scandinavia, pp 1279–1282

11. Davel MH, Martirosian O (2009) Pronunciation dictionary development in resource-scarce environments. In: Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp 2851–2854
12. Davel MH, van Heerden CJ, Barnard E (2012) Validating smartphone-collected speech corpora (accepted for publication). In: Proc. Spoken Language Technologies for Under-resourced Languages (SLTU)
13. Davel MH, Basson WD, van Heerden CJ, Barnard E (2013) NCHLT Dictionaries: Project report. Tech. rep., Multilingual Speech Technologies, North-West University
14. Giwa O, Davel MH (2014) Language identification of individual words with Joint Sequence Models. In: Proc. 15th Annual Conference of the International Speech Communication Association (Interspeech)
15. Giwa O, Davel MH (2015) Text-based language identification of multilingual names. In: Proc. of the Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), pp 166–171
16. Giwa O, Davel MH, Barnard E (2011) A Southern African corpus for multilingual name pronunciation. In: Proc. 22nd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pp 49–53
17. Gustafson J (2009) ONOMASTICA – creating a multi-lingual dictionary of European names. Lund Working Papers in Linguistics 43:66–69
18. van Heerden C, Davel MH, Barnard E (2014) Performance analysis of a multilingual directory enquiries application. In: Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)
19. van Heerden C, Kleynhans N, Davel M (2016) Improving the Lwazi ASR baseline. In: Proc. INTERSPEECH, pp 3534–3538
20. van den Heuvel H, Martens JP, D'hanens K, Konings N (2008) The Autonomata Spoken Names Corpus. In: Proc. 6th conference on International Language Resources and Evaluation (LREC), pp 140–143
21. van den Heuvel H, Réveil B, Martens JP (2009) Pronunciation-based ASR for names. In: Proc. 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp 2959–2962
22. Kamm CA, Shamieh C, Singhal S (1995) Speech recognition issues for directory assistance applications. Speech Communication 17(3):303–311
23. Kgampe M, Davel MH (2010) Consistency of cross-lingual pronunciation of South African personal names. In: 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA 2010), pp 123–127
24. Kgampe M, Davel MH (2011) The predictability of name pronunciation errors in four South African languages. In: Proc. of the 22nd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Emerald Casino and Resort, Vanderbijlpark, South Africa, pp 85–90
25. Llitjós AF, Black AW (2001) Knowledge of language origin improves pronunciation accuracy of proper names. In: 7th European Conference on Speech Communication and Technology (EUROSPEECH), pp 1919–1922
26. Llitjós AF, Black AW, Lenzo K, Rosenfeld R (2001) Improving pronunciation accuracy of proper names with language origin classes. In: Proc. 7th ESSLLI Student Session
27. Loots L, Niesler T (2011) Automatic conversion between pronunciations of different English accents. Speech Communication 53:75–84, DOI 10.1016/j.specom.2010.07.006
28. Maison B, Chen SF, Cohen PS (2003) Pronunciation modeling for names of foreign origin. In: Proc. of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, pp 429–434, DOI 10.1109/ASRU.2003.1318479
29. Modipa T, de Wet F, Davel MH (2009) ASR Performance analysis of an experimental call routing system. In: Proc. 20th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), pp 127–130
30. Modipa TI, Davel MH, de Wet F (2013) Pronunciation modelling of foreign words for Sepedi ASR. In: Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Johannesburg, South Africa, pp 64–69
31. Povey D, Ghoshal A, Boulianne G, Burget L, Glembek O, Goel N, Hannemann M, Motlicek P, Qian Y, Schwarz P, et al (2011) The Kaldi speech recognition toolkit. In: Proc. IEEE 2011 workshop on Automatic Speech Recognition and Understanding (ASRU), Big Island, Hawaii, EPFL-CONF-192584
32. Réveil B, Martens JP, D'Hoore B (2009) How speaker tongue and name source language affect the automatic recognition of spoken names. In: 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp 2971–2974
33. Réveil B, Martens JP, van den Heuvel H (2010) Improving proper name recognition by adding automatically learned pronunciation variants to the lexicon. In: Calzolari N, Choukri K, Maegaard B, Mariani J, Odijk J, Piperidis S, Rosner M, Tapias D (eds) Proc. of the 7th conference on International Language Resources and Evaluation (LREC), pp 2149–2154
34. Réveil B, Martens JP, van den Heuvel H (2012) Improving proper name recognition by means of automatically learned pronunciation variants. Speech Communication 54(3):321–340

35. Schramm H, Rueber B, Kellner A (2000) Strategies for name recognition in automatic directory assistance systems. Speech Communication 31(4):329–338, DOI 10.1016/S0167-6393(99)00066-7
36. Spiegel MF (2003) Proper name pronunciations for speech technology applications. International Journal of Speech Technology 6(4):419–427
37. Strik H, Cucchiarini C (1999) Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication 29(2-4):225–246
38. Thirion JW, Davel MH, Barnard E (2012) Multilingual pronunciations of proper names in a Southern African corpus. In: Proc. 23rd Annual Symposium of the Pattern Recognition Association of South Africa (PRASA), Pretoria, South Africa, pp 102–108
39. Trancoso I, Viana MC (1995) Issues in the pronunciation of proper names: the experience of the Onomastica project. In: Workshop on Integration of Language and Speech, pp 1–16
40. Yang Q, Martens JP, Konings N, van den Heuvel H (2006) Development of a phoneme-to-phoneme (p2p) converter to improve the grapheme-to-phoneme (g2p) conversion of names. In: Proc. 5th International Conference on Language Resources and Evaluation (LREC), pp 287–292
41. Yu D, Ju YC, Wang YY, Zweig G, Acero A (2007) Automated directory assistance system – from theory to practice. In: Proc. 8th Annual Conference of the International Speech Communication Association (INTERSPEECH), pp 2709–2712
42. Zulu PN, Botha G, Barnard E (2008) Orthographic measures of language distances between the official South African languages. Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies 29(1):185–204