

Developing a Multilingual Telephone Based Information System in African Languages

JC Roux^{*}, EC Botha^{**}, and JA du Preez^{***}

^{*}Research Unit for Experimental Phonology, University of Stellenbosch
Private Bag X1 Stellenbosch 7602, South Africa
jcr@maties.sun.ac.za

^{**}Department of Electrical and Electronic Engineering, University of Pretoria
Pretoria, South Africa
botha@ee.up.ac.za

^{***}Department of Electrical and Electronic Engineering, University of Stellenbosch
Stellenbosch, South Africa
dupreez@dsp.sun.ac.za

Abstract

This paper introduces the first project of its kind within the Southern African language engineering context. It focuses on the role of idiosyncratic linguistic and pragmatic features of the different languages concerned and how these features are to be accommodated within (a) the creation of applicable speech corpora and (b) the design of the system at large. An introduction to the multilingual realities of South Africa and its implications for the development of databases is followed by a description of the system and different options that may be implemented in the system.

Introduction

The aims of this presentation are:

- (i) to introduce a project which is probably the first of its kind involving the development of African languages at technological level, and
- (ii) to highlight some language specific traits which will impact on the gathering of speech corpora, as well as on the design of such an interactive multilingual system.
- (iii) To develop a software toolkit which will facilitate future developments in this field.
- (iv) To develop expertise in this field within the South African context.
- (v) To create public awareness of the role of LST applications and to involve potential end-users in participating in future research and development programmes.

South Africa has eleven official languages of which nine are indigenous African languages, i.e. isiZulu, isiXhosa, siSwati, isiNdebele, Sesotho, Setswana, Sepedi, TshiVenda and XiTsonga. The other two official languages are Afrikaans (from Dutch origin) and English. Up to this point in time no real measurable developments in the field of language and speech technology (LST) have taken place in this country. In order to stimulate innovative research and development in various fields the Department of Arts, Culture, Science and Technology (DACST) of the national government set up a special fund, the Innovation Fund to support innovative projects. This paper reports on one of the successful projects entitled *Promoting the development of the official languages of South Africa through language and speech technology applications*. This three year project (working title: *African Speech Technology*) which commenced in January 2000 has five basic aims:

- (i) To develop a prototype multilingual telephone based information system comprising five of the official languages as a demonstration that the indigenous languages of the country may be developed at this level, and that such an application holds tremendous benefits, especially for a country with high illiteracy rates.
- (ii) To build sets of re-usable text and speech resources of these languages for subsequent

In Roux (1998:351) several factors were discussed which were seen to inhibit the development of linguistic resources and speech and language technology applications in a developing country such as South Africa. These factors were mainly of an academic as well as of socio-economic and socio-political nature. Despite this state of affairs there seems to be a growing awareness in various circles that language and speech technology applications have a specific role to play in a multilingual society such as South Africa, and the national government is in the process of setting up a national forum for language and information technology development. Apart from government's involvement in projects of this nature, the newly founded *South African Foundation for Language and Speech Technology Development* endeavours to create an environment conducive for such developments whilst focusing on support from the private sector as possible end-users of these systems.

Figure 1 presents an outline of the macro-architecture of the prototype to be developed. This prototype will be applied as a hotel information and booking system.

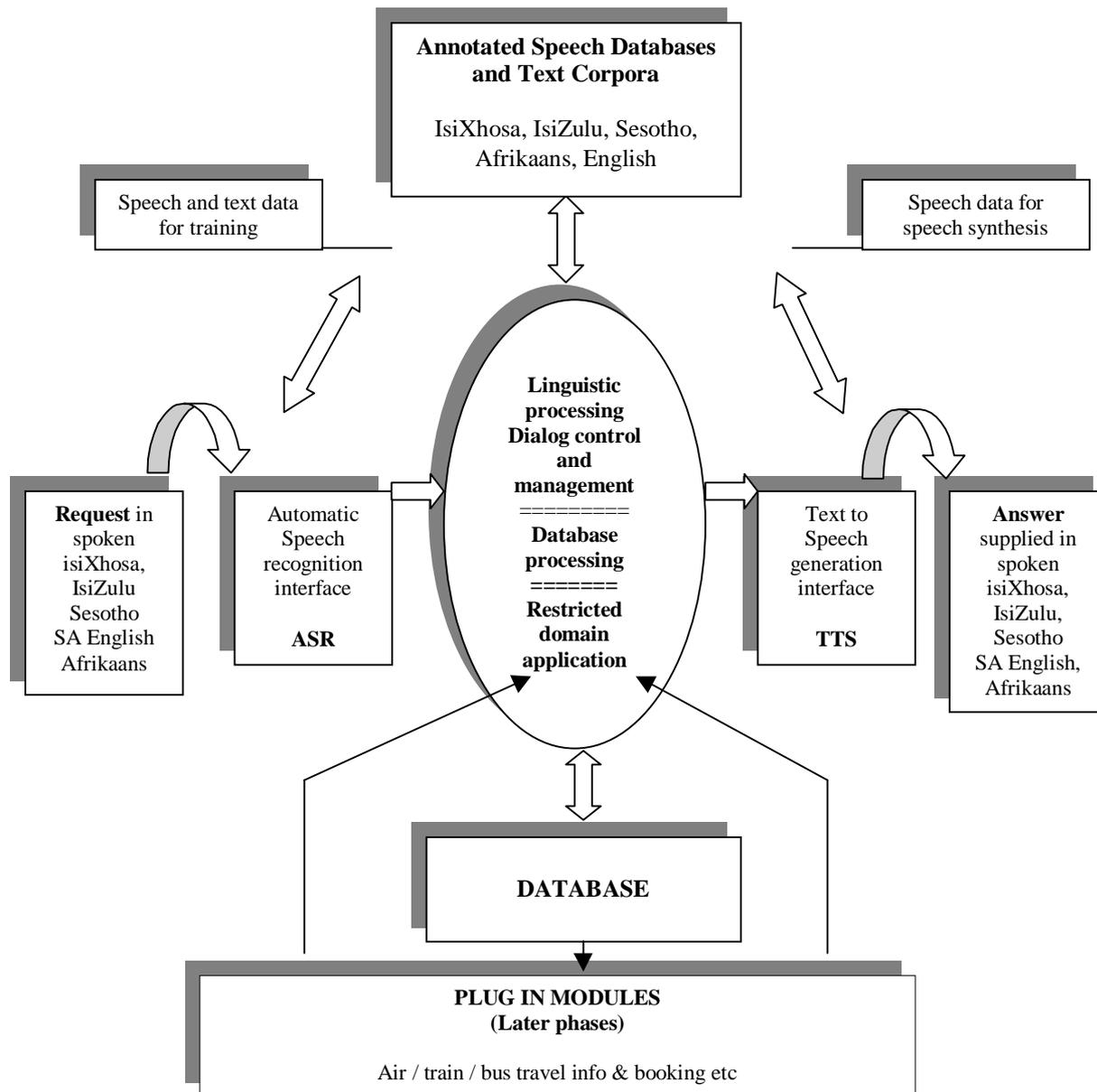


Figure 1: Macro-architecture of an automated multilingual voice based information retrieval and booking system as applicable to the African Speech Technology project.

This classical input-output approach does not represent anything new, however the nature of the languages involved creates a vast number of challenges, especially with regard to particular system design and database construction. These issues will be discussed in more detail in the next sections.

Multilingual Realities in South Africa

There are at least two factors related to multilingual realities in South Africa that influence decisions on the structure of the system to be developed and on the nature of the databases to be acquired.

Home language and language of use

The multilingual complexity of South Africa is clearly reflected in Table 1 which indicates the home language by population group, expressed as a percentage of a total of 40,5 million people. Although these 1996 government statistics (<http://www.statsa.gov.za/census96>) indicate that the first five largest home languages are respectively, isiZulu (22,9%), isiXhosa (17,9%), Afrikaans (14,4%), Sepedi (9,2%) and English (8,6%), the languages of commerce and industry by and large are English and Afrikaans. There is also a clear tendency among black, coloured and Asian speakers to opt for English in these

	African / Black	Coloured	Indian / Asian	White	Unspecified /Other	Total
Afrikaans	0.7	82.1	1.5	58.5	27.7	14.4
English	0.4	16.4	94.4	39.1	23.2	8.6
IsiNdebele	1.9	0.1	0.0	0.1	0.9	1.5
IsiXhosa	23.1	0.3	0.1	0.1	9.9	17.9
IsiZulu	29.5	0.2	0.2	0.1	16.9	22.9
Sepedi	11.9	0.1	0.0	0.0	5.6	9.2
Sesotho	10.0	0.2	0.0	0.0	4.0	7.7
SiSwati	3.3	0.0	0.0	0.0	1.4	2.5
Setswana	10.6	0.4	0.0	0.0	4.4	8.2
Tshivenda	2.8	0.0	0.0	0.0	1.5	2.2
Xitsonga	5.6	0.0	0.0	0.0	2.8	4.4
Other	0.3	0.2	3.7	2.0	1.7	0.6
Total	100.0	100.0	100.0	100.0	100.0	100.0

Table 1: Home language by population group expressed as percentages excluding those who did not specify their language.

circumstances. From a pragmatic perspective (bearing in mind that a person may prefer not to use a home language for the particular transaction), as well as from an “official” perspective (the government is committed to promote the use of the indigenous languages in as many fields of interaction as possible), it is foreseen that the following databases as listed in Table 2 will need to be developed.

Language	Version / Dialect
SA English	1 “Standard” version
	2 Afrikaans version
	3 African / Black version
	4 Coloured version
	5 Indian / Asian version
Afrikaans	1 “Standard” version
	2 Coloured version
	3 African / Black version
isiZulu	1 “Standard” (Dialects excluded)
isiXhosa	1 “Standard” (Dialects excluded)
Sesotho	1 “Standard” (Dialects excluded)

Table 2: Proposed databases

At phonetic level the range of variation is quite large amongst the different languages and language groups. Lanham (1967) identified five distinct varieties of English in South Africa of which SA Black English (BSAE) is one variety. Currently there is a serious debate on whether BSAE is a monolithic entity or whether distinct varieties based on the mother tongue of the speaker may be identified. That is, just as an Afrikaans-English variety may clearly be identified, the question is whether similar varieties such as Zulu-English or Sotho-English do in fact exist. Limited experimental studies have shown that for instance, the rhythm of Afrikaans-English is closer to that of English

L1 speakers than that of Tswana-English (cf. Wissing et al., 2000). Similarly Van Rooij and Grijzenhout (2000) have studied certain idiosyncratic voicing phenomena in Zulu-English that distinguishes it from first language varieties of SA English. It is foreseen that in gathering substantial speech databases this project will make a contribution to the BSAE debate. Prinsloo and Botha (1999) investigated the L1 (Afrikaans mother tongue) and L2 (English mother tongue) pronunciation of the three true diphthongs of Afrikaans and found significant differences in formant space. In another study by Brink and Botha (1999), the acoustic properties of L2 (African language mother tongue) pronunciation of English vowels was compared to their L1 counterparts. In this study, empirical acoustic evidence of the linguistically predicted differences were in fact found. The eventual aim with studies of this kind is to apply these acoustical models to improve automatic speech recognition in Afrikaans and English respectively. From a technical point of view it will be necessary to account for the varieties in Table 2 in a systematic way within the system to avoid a breakdown in communication especially at the speech recognition front end.

Code switching and code mixing

Although it is assumed that a client conversing with a particular system would keep to one particular language, it is a well known fact that within certain communities particularly in South Africa, speakers tend to switch codes, i.e. reverting from one language to the other, or mixing codes, i.e. substituting particular lexical items with counterparts in another language. While code switching is a very personal matter and rather unpredictable, code mixing in the African languages on the other hand seems to be predictable to some extent. That is, mother tongue speakers of any African language will tend to revert to English when referring to *numerals* of any kind, when referring to *time* (be it in a rephonologized form, e.g. “u-half past six” or “u-quarter to one”), and when using the *alphabet*. Multilingual *place names* are also used at random. Many cities in the country have an African language name as well as an English or Afrikaans name.

Johannesburg may for instance be optionally referred to as Gauteng – from Sotho origin, or Egoli – from Nguni origin. Similarly, Pretoria is known as either Tshwane – from Sotho origin or ePitoli – from Nguni (via English) origin.

These realities need to be addressed in developing the above mentioned databases and prompts will have to be set up accordingly to obtain authentic data. As far as system development is concerned specific attention will be paid to accommodate these aspects.

Speech System Design

The proposed architecture of our telephone-based information system to be developed is illustrated in Figure 2. One such a system will be developed for each of the five languages concerned.

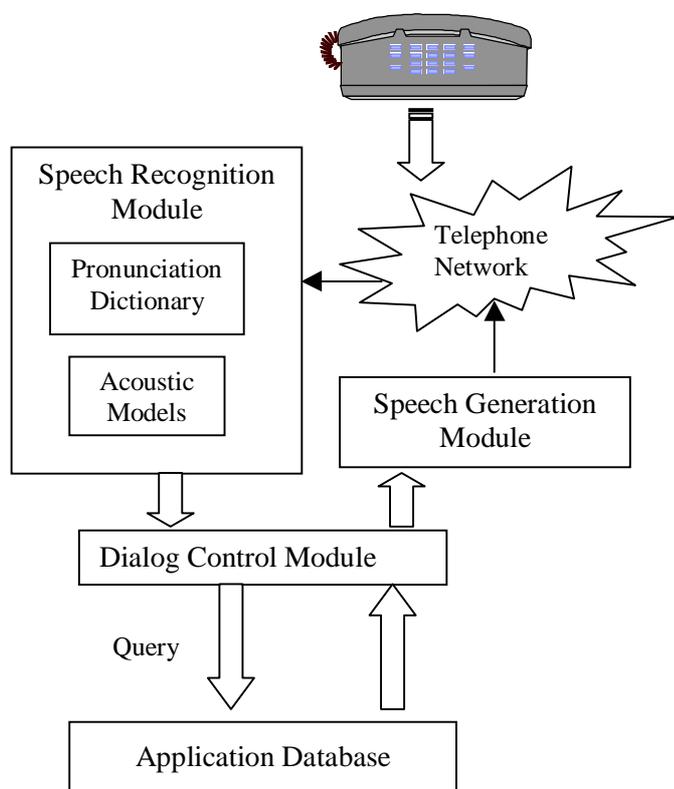


Figure 2: Voice-driven telephone-based information system architecture.

The core of the system is the speech recognition engine, which contains acoustic models and pronunciation dictionaries of the particular language. The recognizer provides the dialog manager module (DMM) with a hypothesis of the content of an unknown utterance together with a confidence measure. The DMM uses three sources of information to generate an appropriate response in the speech generation module (SGM): the recognition hypothesis, with its confidence measure, a dialog model for the particular application and the response from a query to the application database. The query to the database is generated by the DMM, based on the recognition hypothesis. A concatenative word-based model with a generative grammar is envisioned for the SGM. The recognizer gets its input (the unknown utterance) from the caller over the

telephone network and the response (output from the SGM) is sent to the caller through the same network.

Initial language choice

Initially, on receiving the incoming call, it is necessary to activate the specific language-dependent recogniser as required by the caller. This can be done in various ways.

- The simplest approach, and probably the one which we will give initial preference, is to have a separate telephone number for each language available in the system. This makes the languages available explicit before commencement of the transaction, thereby preventing confusion on the caller's side as to whether a specific language is present. To some extent it also circumvents the procedure as to how a language should be activated, thereby allowing the dialogue to more directly proceed to the caller's actual intent. The *role of language specific variation* should, however, not be ignored. For instance, the English versions used in South Africa vary so widely that a single recogniser for all of them is very likely doomed to the outset. In an attempt to avoid these issues, users can be encouraged to stick to their mother tongue (if it is available). It is however debatable whether one can rely on them doing so. On the other hand, advertising telephone numbers for all of the different variants summarised in Table 2, also is somewhat clumsy. Some further possible difficulties to address are the inevitable wrong-language calls, and the speaker switching language during the discourse. The detection of these conditions can be based on estimating the reliability of the ASR, or explicit back-ground monitoring using language/variation identification.
- A next approach would be to query the speaker as to his language preference (possibly using a common language such as English) and recognising the response using speech/word recognition technology. This allows a single front-end/number, but also introduces some extra protocol during the initial stages of the conversation. Variation will, of course, again play an important role. It is also possible (or even likely) that the speaker will ignore this protocol and directly continue in his language of preference. This once again calls for detection of this situation.
- The last option is to make the initial system response in a commonly understood language such as English, possibly with the greeting itself in the various languages available. Automatic language/variation identification can then allow the speaker to proceed in the language of his choice. This approach is currently being used in operator-based services in South Africa, an example being telephone number enquiries. Unfortunately this injects a bias towards a specific (common) language and, if the number of languages being accommodated grows to more than two or three, separate language greetings can become clumsy. The previously mentioned problems of language-switching, of course still remains.

Code switching and mixing

As previously discussed, it is to be expected that callers may, during the call, switch language e.g. from isiXhosa to English or vice-versa. Rather than letting the recogniser fall into disarray due to hypothesizing the contents of speech based on an erroneous language assumption, it will be beneficial to include a *language monitoring module* to allow more elegant recovery from such situations. Alternatively a confidence score describing the estimated reliability in a specific situation, can be used to detect such a situation.

Code mixing, such as using English numbers within African language speech, is also quite common. This probably needs explicit modelling by including it with the vocabulary of the specific language in question.

Language issues specific to the Dialog Model

Possible language specific and culturally bound phenomena may have to be accommodated within the dialog model. In isiXhosa for instance the use of the word *hayi* ("no"), may in some contexts have positive meaning.

Language issues pertaining to ASR

In general, the state-of-the-art ASR systems use mel-scaled cepstral coefficients as features for recognition in either a frame- or segment-based recognition scheme. As the African languages are tone languages (similar to some of the Far Eastern languages where the tone has semantic value), we expect that pitch will have to be included as a feature for reliable recognition of phonemes in our system. In addition, the African languages contain click sounds, which are acoustically similar to the plosives. In initial studies on a discrete-word Xhosa database (Purnell and Botha, 1999), the recognition rates including the clicks seem to be acceptable, but the accompaniments of the clicks (cf. Dogil et al., 1998) were not discriminated sufficiently. No studies on these sounds in natural continuous speech have been conducted. The prominence of clicks may warrant some additional processing for reliable recognition.

Language issues pertaining to Speech Synthesis

All machine response will take place via speech. A variety of approaches towards realising this are possible – all with varying implications as far as multi-linguality goes:

- The first option would be to not synthesize at all. A complete inventory of all possible responses for every likely scenario can be recorded directly. This will, of course, greatly facilitate the naturalness of the response without incurring notable computational burden. Language issues are implicitly handled in this case. However, a little thought will show that this route is only feasible in the most simple of systems. Due to a dependence on earlier information states in the dialog, the required number of responses typically increases combinatorially with the number of concepts in the system.
- At the other extreme text-to-phoneme conversion (Damper et al., 1999) combined with articulatory

and/or formant synthesizers (Parthasarathy and Coker, 1992) can result in very flexible text-to-speech (TTS) synthesis completely driven from the textual versions of the desired responses. In principle therefore, the inventory of possible responses need not even be known beforehand. On the downside, however, it is hard to attain good natural sounding speech with such a system. This approach must take the target language fully into account. This includes phonemic, phonetic, phonological and prosodic effects.

- An approach combining aspects of both of the previous two, is to concatenate recorded speech sub-units the range of which can range from sub-phones (Moulines and Charpentier, 1990) through words to sub phrases. The smaller these sub-units, the greater the flexibility but the more severe the lack of naturalness and vice versa. Also, with very small units, coarticulation effects become prominent and require special counter-measures (Donovan and Woodland, 1999).

From these options we have judged the concatenation of words a suitable choice. Since this avoids the need for synthesizing from scratch and also includes within-word co-articulation effects, this simplifies the synthesis system considerably. As in all concatenative synthesizers, facilities are needed to eliminate/reduce discontinuities, such as relative energy levels, coarticulation effects and pitch, at the concatenated edges. The following language-dependent aspects also need attention:

- The synthesis vocabulary will have to be matched to the application on a per-language basis. For new applications it is likely that it will have to be extended with new recordings - this is not seen as a major obstacle. Although unlikely to be part of the initial system, some interesting alternatives to merely making new recordings, are also possible. Using good quality voice conversion (Moulines and Sagisaka, 1995) facilities, it may even be ultimately possible to draw the required word recordings directly from the speech databases required for the ASR work.
- The prosody of the whole speech segment should be natural for a mother-tongue speaker of the language in question. Under this we include, amongst others, global pitch pattern, inflection, stress placement and duration of the various sub-units. Some of these are to a certain extent also interdependent.

All of these will be highly dependent on the nature of the specific phrase and/or language in question. Foremost then, this implies language-specific knowledge as to how these parameters vary according to the type and content of the speech. It also implies the ability to locate in these pre-recorded words the segments of interest, so that they can be manipulated. If this is done via aligned transcriptions, expanding the range of synthesized responses necessarily will be a more expensive process. Alternatively, software will be required to automatically locate these segments.

Conclusion

Developing this system obviously poses a number of challenges, especially if it is taken into account that there is no real history on similar types of developments within the context of the African languages. It is challenging enough to develop a monolingual system in this fashion and we have no illusions on the magnitude of this venture. Fortunately, however, any "late starter" has the benefit of previous experiences of colleagues in the field in general. It is believed that this project will not only introduce African languages in a meaningful way to this exciting field, but that it will also make some contribution towards the "African Renaissance."

References

- Brink, J.D. & Botha, E.C. (1999). Towards an acoustic comparison of first and second language South African English. In Proceedings of the Tenth South African Workshop on Pattern Recognition (paper #0009). PRASA, Stellenbosch, South Africa.
- Damper, R.Y., Marchand, M.A. & Gustafson, K. (1999). Evaluating the pronunciation component of text-to-speech systems for English: a performance comparison of different approaches. *Computer Speech and Language*, 13 (2): 155-176.
- Dogil, G, Mayer, J & Roux, JC. (1998). Syllables and unencoded speech: Clicks and their accompaniments in Xhosa. In J Rennison & K Kuhnhammer (Eds.) *PHONOLOGICA* 96: Syllables?! (pp. 49-60). The Hague: Holland Academic Graphics.
- Donovan, R. & Woodland, P. (1999). A hidden Markov-model-based trainable speech synthesizer. *Computer Speech and Language*, 13 (3): 223-241.
- Lanham, L.W. (1967). *The pronunciation of South African English*. Cape Town: A.A. Balkema.
- Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing using diphones. *Speech Communication*, 9:453-468.
- Moulines, E. and Sagisaka, Y. (1995). Voice conversion: State of the art and perspectives (guest eds.). *Speech Communication*, 16 (2).
- Parthasarathy, S and Coker, C. (1992). Automatic estimation of articulatory parameters. *Computer Speech and Language*, 6:36-75.
- Roux, JC (1998). SASPEECH: Establishing speech resources for the indigenous languages of South Africa. In A Rubio et al (Eds.), Proceedings of the First International Conference on Language Resources and Evaluation. Granada, Spain. European Language Resources Association – ELRA. Vol 1, 343-350.
- Prinsloo, C.P. & Botha, E.C. (1999). Towards modelling the acoustic differences between the diphthongs of Afrikaans and English L1 and L2 speech. In Proceedings of IEEE Africon, Vol 1 (pp. 207-210). Cape Town, South Africa.
- Purnell, D.W. & Botha, E.C. (1998). Automatic phone segmentation and labelling of Xhosa data. In Proceedings of the Tenth South African Workshop on Pattern Recognition (pp 23-28). PRASA, Stellenbosch, South Africa.
- Van Rooij, B & Grijzenhout, J (2000). Voicing Phenomena in Zulu-English. In Proceedings of Workshop on Black South African English,

International Conference on Linguistics in South Africa (pp. 81-90). Cape Town.

- Wissing, D, Gustafson, K, & Coetzee, A (2000). Temporal organisation in some varieties of South African English: Syllable compression effects in different types of foot structures. In Proceedings of Workshop on Black South African English, International Conference on Linguistics in South Africa (pp. 59-69). Cape Town.

Acknowledgment

This project is supported by the Innovation Fund (project number 21213) of the Department of Arts Culture Science and Technology of the South African National Government.