Problems and solutions in African tone language Text–To–Speech

Dafydd Gibbon, Eno–Abasi Urua, Moses Ekpenyong

Universität Bielefeld, Germany; University of Uyo

gibbon@uni-bielefeld.de; anemandinyene@yahoo.com; ekpenyong_moses@yahoo.com

Abstract

One of the most useful HLT systems for use with non-literate communities is Text–to–Speech, with typical applications in health, market and education information dissemination. However, current TTS development environments are far from being language independent: adaptation of an existing voice in one language to a voice in another is a common procedure. During development of a TTS prototype for Ibibio, a Nigerian Niger–Congo, Lower Cross language, in the *Local Language Speech Technology Initiative*, several levels of infrastructural and linguistic problems were identified. This contribution outlines these problems, concentrating mainly on requirements for African tone language TTS, and formulates solution strategies for both rule–driven and data–driven TTS development.

1. Introduction

One of the most useful HLT systems for use with non-literate communities is Text–to–Speech, with typical applications in health, market and educational information dissemination. However, current TTS development environments are far from being language independent: adaptation of an existing voice in one language to a voice in another is a common procedure.

In the LLSTI project [1], the adaptation procedure was applied to a Nigerian tone language, Ibibio (ISO 693-2: nic; Ethnologue: IBB), the official language of Akwa Ibom State in the Nigerian Federation. Adaptation is plausible when languages are prosodically and phonemically similar but severe problems arise when languages are very dissimilar (e.g. 'intonation languages', for which TTS systems are typically developed, vs. 'tone languages', e.g. Ibibio).

These problems can be generalised to other African tone languages, which have lexical (phonemic) tone but, unlike Asian tone languages, also morphemic tone: e.g. in Ibibio near/far tense is marked by LH/HL tones on tense morphemes. Tone-morpheme combinations could be used in unit selection, but the problem is compounded by the lack of orthographic tone marking, making morphological tone assignment effectively an 'AI complete' problem.

Additionally, 'terraced tone' patterning generated by automatic and non-automatic downstep requires a clean interface for signal processing of pitch: pitch has a 'hard' semantic function rather than a 'soft' pragmatic function in intonation languages [2].

On the language side, agglutinative inflectional morphology generally makes it infeasible to list all inflected word forms except for small vocabularies and complex subject-verb-object person concord makes morphological tone assignment difficult.

Finally, specific problems of sparseness and text inconsistency in written but non-standardised languages arise.

The present contribution outlines the initial developmeng issues involved in preparing an experimental TTS prototype for Ibibio, and presents detailed computational solutions for three areas: a tone model; a model for non-local morphological dependencies; corpus preparation and processing for a tone language with a small and inconsistent text base.

2. Overview of issues

The development of an experimental TTS prototype for a new language, from project design through resource acquisition to system design, implementation and evaluation is a long haul, even with older technologies using diphone TTS shells such as MBROLA [3] or more modern technologies with unit selection shells such as Festival [4] or BOSS [5].

2.1. Linguistic issues: language typology

On the one hand, much depends on the typology of the language concerned, both from the perspective of the creation of a unit database resource for the TTS system, and from the perspective of the processing of text input to the TTS system. Key issues include:

1. Syllable phonotactics. The complexity of syllable phonotactics is a major determinant of the size of the unit database resource: whether a language is fundamentally CCCVVCC, like many Indo-European languages, or CV, CCV, CCVC, like many West African languages, leads to a significant difference in combinatoric options, and therefor in resource size. Syllable phonotactics includes prosody: syllables may be long or short, open or closed, strong or weak, stressed/accented or unstressed/unaccented, tonal or non–tonal, depending on the language.

2. Inflectional morphotactics. Inflectional morphotactics determines, for example, whether table lookup or rule–based techniques are most appropriate for handling the vocabulary and its adaptation to grammatical context: English tends towards the isolating type, with little inflection; other Indo-European languages are fusional, tendentially with single inflectional suffixes, each with complex grammatical meanings; other languages, such as Finno-Ugric languages and many African languages, are agglutinating, with chains of suffixes, each of which has a simple grammatical meaning. Inflectional morphotactics also includes prosody: inflection may affect stress patterns, as in the German singular noun *DOKtor*, plural *DokTORen*, or tonal patterns, as in Ibibio verbal morphology where, simplifying slightlym, rising and falling tones on a future or past prefix particle represent distal

future or past, and proximal future or past, respectively (yàá, yáà; màá, máà).

3. Morphotactics of word formation. Derivation with one root and affixes, compounding with more than one root, determine the composition of the vocabulary. Prosody also plays a central role: in languages such as English, stress patterns (IMport/imPORT; TELEphone/teLEphony/telePHONic; BLACKboard) are involved. In West African language, low or high tonal interfixes may be involved, as in Ibibio: ènò (gift) +´ (tonal interfix) + àbàsì (God) = ènòábàsì (personal proper name).

4. Sentence structure. The combinatorics of words into larger units takes place in conjunction with inflectional morphology: different orders of subject, object and verb in simple sentences; different concord conditions in simple sentences (e.g. subject–verb, object–verb, ergative); serial verb constructions; different conventions for embedded sentences.

2.2. Resource issues

Many aspects of TTS development depend on the development environment: from data resources (corpora, lexica, grammars), through hardware and software, to the instutional environment, with its personnel structures and economic basis, and in this respect conditions vary from region to region, and from continent to continent.

In relation to Europe, resource conditions differ significantly in all of these respects in West Africa, and, of course, the issue of speech synthesis for local African languages is itself very heterogeneous.

The experimental prototype development project for Ibibio was conducted at the University of Uyo, located in Uyo, the capital of the Akwa Ibom State in the Nigerian Federation, in the South East of Nigeria, between the Cross and Calabar rivers. Universität Bielefeld was coordinating cooperation partner in the project. The official language of Akwa Ibom State is Ibibio. The 2 million speakers of Ibibio constitute a rather large application pool for speech synthesis applications in education, health–care, agricultural and geographical information systems. The Department of Linguistics and Nigerian Languages at Uyo University is already extensively involved in AIDS/HIV education in the mass media, and well–informed and highly motivated in respect of other applications of information technologies in such fields.

The particular issues to be addressed in the Ibibio TTS project emerged only slowly during the project itself. The main factors were:

1. Human resources. Education and training goals of personnel, both in linguistics and in computer science, were different, because of different local requirements in vocational training, and in different scientific traditions. Training on language documentation techniques were needed by the Uyo partners, and training on the Ibibio language was needed by the Bielefeld partners.

2. Empirical resources. Developing a TTS system is clearly a maximally hard task for an unwritten language, understating matters slightly. A language like Ibibio, whose writing system is only two or three decades old, is in an intermediate position between an unwritten language and a language with an old-established orthography:

   - For historical reasons there are competing orthographies: they differ in details.
   - Existing documents created by untrained personnel tend to be highly inconsistent in orthography, punctuation and format (though this is hardly different from documents in many other languages the world over).
   - The body of available texts is small, for IT purposes the available data are therefore very sparse.
   - Available lexicographic material is also very restricted, and in a legacy paper format.
   - Available information on grammar is neither precise enough nor complete enough to be of much use for automatic processing (e.g. parser development), though the morphology is more complete, and the phonology is in fact complete.

3. Infrastructural resources. The ubiquity of electricity and of internet connectivity which academic researchers expect in, for example, European contexts, is either not given, or very expensive. Power cuts can be frequent, and are sometimes predictable due to known times when demand is likely to outstrip supply. For most hands-on computational work either a desktop with backup generator or a laptop with extra batteries is required. The usual solution is a laptop. Unsealed magnetic storage media (floppy disks) have a very short life, because of dust, heat and humidity.

The last of these points is crucial for local language IT development: creaming off development work into well–equipped labs in more northerly (or southerly) latitudes is likely to fan the flames of reputation for the labs concerned, but is not likely to lead to significant value for or acceptance by users of the local language unless local developers are responsible for development on an equal footing; an optimal solution would be partnerships between well–equipped labs and regional, less well-equipped labs.

3. A note on tonal typology

In this and the following sections, attention is focussed on the consequences of differences in language typology for TTS development, in the present case on the typology of tone [6], [7].

The uses of pitch in the languages of the world are very varied, ranging from phonemic functions in languages such as Mandarin, through morphemic and morphosyntactic as in the Niger-Congo and Bantu languages of Africa, to sentential functions in Indo-European languages, and discoursal intonation functions, possibly in all languages.

A rank–based typology of pitch functionality is given in Table 1. Pitch functionality at the phonological and morphological ranks (which determine word prosody) is usually referred to as *tone*, and at the syntactic, text and dialogue levels (which determine discourse prosody) as *intonation*.

3.1. Formal preliminaries

It is known that—at least for the organisation of forms—the basic structure of phonological and prosodic systems can be modelled by regular (linear) languages and regular (linear) grammars (equivalently: finite state automata, FSAs). Finite State (FS) modelling holds even where hierarchies are involved in the organisation of pitch systems: the hierarchies put forward

Table 1: Pitch–functional rank levels.

Phonology	phonemic 'lexical tone'
Morphology	morphemic sub–rank: - morphophonemic 'lexical tone' derived word sub–rank: - morphological templatic tone: compound word sub–rank: - tonal interfixation inflectional sub–rank: - morphosyntax
Syntax	phrase sub–rank: - templatic phrasal intonation sentence sub–rank: - sentence intonation: phrasing, accentuation, nucleus
Text	textual 'paragraph' intonation: cohesive pitch contours, focal and contrastive accentuation
Dialogue	dialogue control, emotion

so far are either of finite depth or are purely right–branching or left–branching (but cf. [8]) and thus formally FS–equivalent.

Explicit applications of FSAs to modelling intonation forms have been available since [9], the 1970s IPO model [10], [11] and [12]. Implicitly, many other intonation models, including [13], are also FS models. In this paper, FS intonation patterning will not be dealt with further.

FST modelling of the tone–phonetics interface started with a model of two Niger–Congo languages (Baule, Kwa; Tem, Gur) in [14] (cf. also [2]). The technique was extended to Mandarin tonal sandhi, mapping lexical tone sequences to other lexical tone sequences, in [15].

3.2. Morphotonemic–phonetic interface (tone)

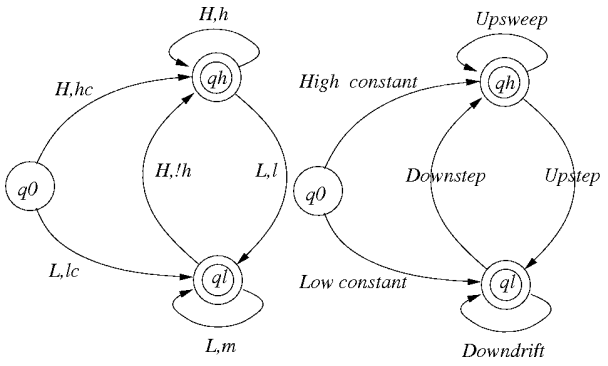

Figure 1: (a) Basic 2–tone Niger-Congo FST; (b) Generalisation of tone FST mapping types ([2]).

The complex tonal structure and functionality of Western and Central African languages has often appeared in the literature: tone terracing vs. discrete level tone patterns, automatic and lexical downstep, upstep, downdrift and upsweep, tonal blocking, tone–depressor consonants. Of specific interest here is tone terracing. Tone terracing is a phonemic-phonetic mapping, and was modelled in Metrical Phonology by right–branching trees [16]. Since right–branching trees are accessible to FS modelling, [14] concluded that tone terrace mapping is formalisable with FSTs, and provided FST models of terracing in Baule (Kwa, Ivory Coast; cf. Figure 2(a) for a locally non–deterministic FST with 1-place lookahead) and in Tem (Gur, Togo). In [2] these models were generalised to a schema for any

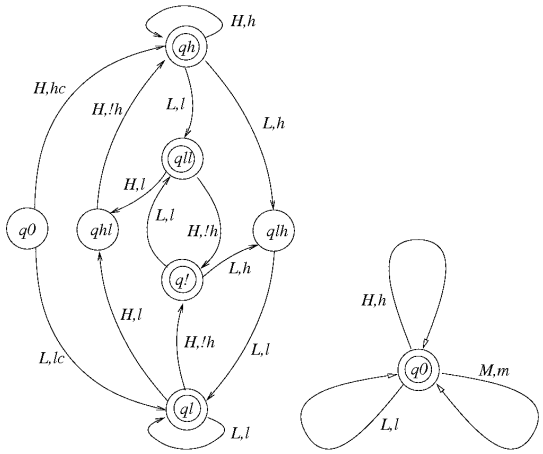

Figure 2: Variants: Baule FST (with lookahead) and 3-tone FST.

two–tone terraced system; cf. Figure 1(b). Figure 2(b) shows the simple FST model required for 3–tone discrete–level tone systems.

Formally, the basic two–tone FST is the union of just two simpler FSTs, one starting with high tones and one with low tones. The simpler FSTs are isomorphic but for the labelling.

The topology of the African tone FSTs shown in Figures 1 and 2 has very general and symmetrical properties, and is thus quite different from the more idiosyncratic topology of the Mandarin tone FST provided by [15], in addition to the difference in levels of representation. It should be noted that this kind of explicit tone grammar modelling has not yet found its way into tone description in 'mainstream phonology' (cf. the Handbook contributions of Odden and Yip in [17], and many later publications).

3.3. Phonetic–acoustic interface (tone)

For modelling the pitch time function for high and low tone sequences, an asymptotic function similar to that of [18] was used (the reference unit for pitch association is the syllable):

$$pitch_{i+1} = tone * (pitch_i - baseline) + baseline$$

with speaker–specific initial, $baseline$ and $tone$ values, where $tone$<1 for low tones, $tone$>1 for high tones. The model is linear and local. Using the FSTs shown in Figures 1 and 2 it is straightforward to implement a transducer in which the phonetic output symbols are replaced with the appropriately instantiated numerical functions. A 'toy instantiation' can be illustrated informally (not in the detail required for Ibibio) as follows:

High tone factor:	1.1
Low tone factor:	0.8
Downstep factor:	1.3
Upstep factor:	0.6
Baseline component	100 Hz
Initial high component:	80 Hz
Initial low component:	80 Hz
Input (tones):	H L H L
Output (Hz):	180 148 162 137

The 'real world' empirical basis for the actual Ibibio model was induced by an exhaustive prosodic data mining algorithm applied to Ibibio data [19].

4. Morphophonemic and morphosyntactic tone

The previous section dealt with form–form interfaces; the present section deals with morphotonology at the syntax–morphology interface. Such factors are frequently referred to in passing in the literature, but, like metrical trees, never provided with a grammar model. The factors involved in Ibibio morphotonology (simplifying for brevity of presentation) are:

1. Part of Speech (POS): there are four tonological categories determined by POS in Ibibio [20, 21]:
   (a) Nouns: lexical tone with phonemic functionality, comparable with tone in East Asian languages; òbû 'crayfish' - óbû 'dust'.
   (b) Verbs: Fixed tonal templates, modifiable by inflexion and verb subcategorisation.
   (c) Autonomous tonal function morphemes: HL meaning 'proximate future/past' and LH meaning 'non–proximate future/past', with the tense prefixes *yaa* and *maa*, respectively, e.g. n-yaa-ka 'I will go (sometime)'
   (d) Composition template morphemes: in word–formation, superimposed patterns which function as 'interfixes': *èno* 'gift', *àbàsì* 'God' form a compound:
   ènò + high Tone + *àbàsì* → *ènòábàsì*
   (cf. the interfix function in German: *Liebe<u>s</u>brief*).
   (e) Templatic function words and affixes: determiners, which are NP-initial, tend to have the same pattern, high–low, while quantifiers, which are NP-final, tend to have the opposite pattern, low–high.

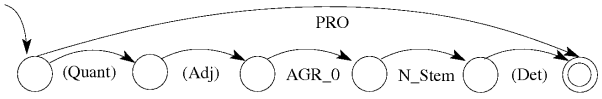

Figure 3: Ibibio Noun Phrase FST.

The autonomous tonal function morphemes and the tones of templatic function words and affixes are particularly interesting from the FS processing point of view: the sub–sentential structures which they mark are non–hierarchical, and can be modelled with FS devices; cf. Figure 3. The symbols in the morphosyntactic FST stand for pairs <POS,TONES> of a part of speech (POS) and the appropriate set of tones for each POS. Consequently, a full description of tonal morphosyntax requires three aligned levels of representation (tiers, tapes). The individual components have been implemented for test purposes, but the full architecture has not yet been implemented. For this, two models are being considered: Kay's multi–tape FSTs for Arabic templatic morphology [22], or a cascade of a morphosyntactic FST as in Figure 3 and a terracing FST as in Figure 1, which can be composed for efficiency into a single FST [23].

5. Solutions — a first approach

Solutions to the infrastructural are, fundamentally, questions of economics, cooperativity and training, and are too general to be amenable to useful discussion here. However, solutions for the linguistic problems in the TTS context are feasible. These problems and possible solutions to them are detailed here.

First, the general solution to language modelling lies in the development of Finite State Transuders:

1. FSTs can handle syntactic and morphosyntactic structures within the domain of simple sentences and right–branching serial constructions, or for morphosyntax, including subject–verb–object concord.
2. FSTs can handle the compositional properties of syllables (trivially, since the set of syllables is fininte).
3. FSTs can handle the phonetic interpretation of tone sequences (tone sandhi) in terms of tone terracing, automatic downstep, upstep, upsweep, downdrift, and related tone sequencing patterns.

These principles have been applied in the creation of test generators of sentences and tone patterns, in order to perform extensive checks of well–formedness of linguistic descriptions. For example, in order to increase the precision of available information on grammatical topics, a multi–tape finite state sentence generator was implemented and used for formulating and testing precise hypotheses on Ibibio syntax. To check the validity of tone terracing, an acoustic generator was implemented.

It is not immediately clear, however, how to utilise this knowledge in the context of a statistical unit–selection TTS system. Two main issues arise in connection with the incorporation of tones:

1. Offline: How are the speech corpora to be annotated in order to include the characteristic effects of tone terracing into the unit selection process?
2. Runtime: Text preprocessing requires annotation of tones. Ibibio orthogaphy does not provide tone marking — why should native speakers need it? In general, native speakers can easily cope with the lexical ambiguity resulting from the lack of tone marks, just as lexical ambiguity in other languages resulting from other causes can easily be handled.

The actual solution in both these cases does not become clear until a more abstract position is taken: essentially, the tone sequencing FSTs simply define effects at adjacent positions in an utterance, starting at the beginning.

Consequently, the required corpus annotator (human or automatic) will assign lexical and morphosyntactic tones, and these will be assigned a positional number, e.g. a sequence of High (H) and Low (L) tones such as "H H L H L" will be assigned "H_1 H_2 L_3 H_4 H_5", where the indexed annotations are those which are relevant for defining units.

The text markup generator will do exactly the same, *mutatis mutandis*: first, lexical tones are assigned from a dictionary, and morphosyntactic tones are assigned either on the basis of the morphological FST, or if ambiguities remain, on the basis of context–sensitive default assumptions (e.g. a L tone or a HL tone sequence).

6. Conclusion

The form, structure and function of the tonal systems of Ibibio differ fundamentally from those of Indo–European languages and from East Asian tone languages. Ibibio was selected as a typical Niger–Congo language, with a wide range of morphotonological features which are known to be characteristic of these languages.

Two main conclusions can be drawn. First, the tone system of Ibibio is better classified as morphophonemic and morphotactic tone system than simply as a phonemic 'lexical tone' system. Second, because of this, the language has a typologically very different use of pitch from from the languages which have so far been modelled in TTS systems. Consequently, it is not possible to apply results of work in accent–intonation languages or phonemic lexical tone languages blindly to applications in speech technology for Ibibio and other Niger-Congo languages in West, Central and Southern Africa.

For applications in speech synthesis, the solutions outlined in this study solve at least some of the problems involved in for both resource and system development for African language TTS:

1. Resource development: in corpus design and markup for unit selection (whether at diphone or higher unit levels) the tone–relevant morphophonemic and morphosyntactic contexts have to be included.
2. System development: in the linguistic TTS component, appropriate tone–oriented parsing and tagging is required as a basis for unit selection search and costing.

The solutions outlined here are linguistic results. The engineering application of these results in the *Local Languages Speech Technology Initiative (LLSTI)* project and the evaluation of these results are described in [1].

7. References

[1] Tucker, Roger & Shalonova, Ksenia (2005). Supporting the creation of TTS for local language voice information systems, INTERSPEECH-2005, pp. 453-456.

[2] Gibbon, Dafydd (2001). Finite State Prosodic Analysis of African Corpus Resources. 7th EUROSPEECH Conference, Aalborg, Denmark, pp. 83–86.

[3] Dutoit, Thierry (1998). *An Introduction to Text–To–Speech Synthesis*. Dordrecht, etc.: Kluwer Academic Publishers.

[4] Taylor, Paul, Alan Black & Richard Caley (1998). The architecture of the Festival Speech Synthesis System. 3rd ESCA Workshop on Speech Synthesis, pp. 147-151, Jenolan Caves, Australia.

[5] Klabbers, Esther, Karlheinz Stöber; Raymond Veldhuis, Petra Wagner & Stefan Breuer (2001): Speech synthesis development made easy: The Bonn Open Synthesis System in: Proceedings of Eurospeech. Aalborg.

[6] Gut, Ulrike & Dafydd Gibbon, eds. (2002). *Typology of African Prosodic Systems 2002*. Bielefeld Occasional Papers in Typology 1. Universität Bielefeld.

[7] . Gibbon, Dafydd & Eno–Abasi Urua (2006). Computational morphotonology in Niger-Congo languages. Proceedings of Speech Prosody 2006, Dresden, Germany.

[8] Steedman, Mark (1991). Syntax, Intonation and 'Focus', (1991) In Ewan Klein and Frank Veltman, (eds.), *Natural Language and Speech*. Mahwah, N.J.: Lawrece Erlbaum Associates, pp. 331-342.

[9] Reich, Peter A. (1969). The finiteness of natural language. *Language*, 45, 831-843.

[10] 't Hart, Johan & Antonie Cohen (1973). Intonation by rule, a perceptual quest. *Journal of Phonetics* 1, p. 309–327.

[11] Pierrehumbert, Janet (1980). The phonology and phonetics of English intonation. Diss. Massachusetts Institut of Technology.

[12] Gibbon, Dafydd (1981). A new look at intonation syntax and semantics. In A. James & P. Westney, eds., *New Linguistic Impulses in Foreign Language Teaching*. Tbingen, Narr, pp. 71-98.

[13] Hiroya Fujisaki (1988). A note on the physiological and physical basis for the phrase and accent components in the voice fundamental frequency contour. In Osamu Fujimura, ed., *Vocal physiology: Voice production, mechanisms and functions*, pp. 347-355. New York: Raven.

[14] Gibbon, Dafydd (1987). Finite State Processing of Tone Systems. *Proceedings of the European Chapter of the ACL*, Copenhagen, pp. 291–297.

[15] Jansche, Martin (1998). A Two-level Take on Tianjin Tone. In: Gert–Jan Kruijff & Ivana Kruijff–Korbayová, eds., *Proceedings of the Third ESSLLI Student Session*, 10th European Summer School on Logic, Language and Information, Saarbrücken, Germany, pp. 162–174.

[16] Clements , G. N. (1981). The hierarchical representation of tone features. *Harvard Studies in Phonology* 2. Indiana University Linguistics Club.

[17] Goldsmith, John A., ed. (1995). The Handbook of Phonological Theory. Oxford: Blackwell.

[18] Liberman, Mark & Janet Pierrehumbert (1984). Intonational invariance under changes in pitch range and length." In: Mark Aronoff and Richard T. Oehrle (eds.), *Language Sound Structure*. Cambridge: MIT Press. pp. 157-233.

[19] Gibbon, Dafydd, Eno–Abasi Urua & Ulrike Gut (2003). A computational model of low tones in Ibibio. *Proceedings of the International Congress of Phonetic Sciences*, Barcelona, 2003, pp. 623-626.

[20] Essien, Okon E. (1990). *A Grammar of the Ibibio Language*. Ibadan: University Press Limited.

[21] Urua, Eno–Abasi (2000). *Ibibio Phonetics and Phonology*. Cape Town: Centre for Advanced Studies of African Society.

[22] Kay, Martin (1987). Nonconcatenative Finite-State Morphology. *Proceedings of the European ACL Conference*, Copenhagen, pp. 2-10

[23] Kaplan, Ronald M. and Martin Kay. 1994. Regular Models of Phonological Rule Systems. *Computational Linguistics*, 20:3, pp 331-378.

8. Acknowledgments

For much discussion and joint work on problems of tonal typology we are also heavily indebted to Shu-Chuan Tseng, Academia Sinica, Taipei, Taiwan, to Firmin Ahoua, Université de Cocody, Abidjan, Côte d'Ivoire, and to Etienne Barnard and his team at CSIR, Pretoria, South Africa.