RESEARCH UNIT FOR EXPERIMENTAL PHONOLOGY

UNIVERSITY OF STELLENBOSCH (RUEPUS)

ANNUAL REPORT: 1999/2000

1 BACKGROUND

This progress report mainly covers activities of the **second year** of the second cycle of the Unit, i.e. activities that have taken place in the financial year ending **March 2000**. As with the report of 1999 there is a measure of overlap in reported activities present in this report. In order, however, to maintain a wider perspective with respect to past and present (third year) activities that will have a direct bearing on the **fourth and last year** of RUEPUS (April 2001-March 2002), it is necessary to provide some information on activities extending up to September 2000. The report will be presented in the prescribed format for annual progress reports as specified by the NRF.

Following specific suggestions presented to and accepted by the Advisory Board at its annual meetings in 1997, 1998 and 1999, the focus of RUEPUS has been redirected towards research and development in the field of language and speech technology applications with specific reference to the indigenous languages of South Africa. (Please consult previous reports in this regard). This led to the acceptance of a new strategic plan in 1998 that is taken into account in this report.

1.1 Mission and vision

The **mission** of RUEPUS remains unchanged as described in the 1997/8 report, i.e. as the active promotion of the official languages of South Africa at **technological level** through research, facilitation and development activities with a focus on the African languages and the speakers of these languages.

The **vision** upheld by RUEPUS is that all languages spoken in South Africa, and more specifically the African languages, be *developed at technological level* to such an extent that

- They will meet the challenges of a technologically driven new millennium;
- A *new generation of language engineers* will be created which will play a central role in the development of language based applications which will infiltrate all walks of life in the new century;
- A technology driven *national resource facility* making provision for speech and text corpora will be developed and maintained.

The last vision mentioned here is proving to become one of the main foci in deciding on the future of RUEPUS after the expected termination of financial support by the NRF in 2002. (See section 4 below).

1.2 Long term goals

The 1998/1999 Report envisaged the following long-term goals:

"The strategic plan discussed and accepted by the Advisory Board in 1998 envisages that through focused activities RUEPUS will be functional in setting up a new, financially self sustaining institution with exclusive focus on **language engineering** (LE) within the African languages. RUEPUS should eventually become part of this new institution in the final year of its second cycle (i.e. in 2002). This new centre, the **(South) African Language Engineering Centre ((S)ALEC),** will operate on a national scale to the benefit of all potential researchers in LE in the (South) African languages bringing together

resources from the academic community with financial input from business and industry and government according to a model similar to that implemented by the European Union."

In principle this goal is still pursued, however, with the following adaptations:

i. That the proposed name be changed to South African Language and Speech Resource Centre (SALASREC).

ii. That the possibility be investigated that SALASREC be positioned at the University of Stellenbosch as a virtual centre managed by a semi-private company possibly within the UniStel Holdings concept.

(Motivation to be presented in section 4 below.)

2 PROGRESS REPORT

2.1 Achievement of goals

The goals set by RUEPUS reflect on

- **Academic activities** (the running of research projects and the dissemination of research results through publications and conference presentations),
- **Networking** (making contact with colleagues at national and international level, *inter alia* by organising conferences/seminars and/or participating in other conferences/seminars)
- **Research capacity building** (at graduate and post graduate levels as part of formal academic courses; through dedicated research capacity building projects; through specialised workshops)
- **Fund raising** to create a self-sustainable infrastructure conducive to research and development.

In assessing the achievement of the **academic aims** it may be stated that following a very successful output the previous year, the quantity output with respect to publication in accredited journals was lower this year. The reasons for this are

i. The timing of publications - two major articles have been presented for international publication, however, the turnover time for publication is substantially longer than that for local publications; two further articles presented at national level are to be published later this year of early next year,

ii. A deliberate focus on participation in international conferences which allows for relatively swift feedback on particular issues, especially in a field where technology is under constant change and where it is important to keep abreast with the state of the art developments,

iii. The amount of time spent on issues related to language policy and technology where National Government had to be convinced of the necessity of the development of Human Language Technologies in South Africa, in order to open new avenues for research and development for linguists, computer scientists and electronic engineers. (See 2.3.3.3)

As far as **networking** is concerned the overall visibility of RUEPUS at national as well as at international level was maintained through conference participation and international contacts. International research and development co-operation was strengthened with the Director being elected in 1999/2000 to the editorial boards of, respectively, *Afrika und Übersee* (Hamburg), the *Journal of the International Phonetic Association* (JIPA) (Los Angeles/London), and as a member of the *International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques* (COCOSDA).

As far as **research capacity building (RCB)** is concerned no new actions are to be reported save the completion or near completion of advanced studies for degree purposes. (Reports in the appropriate sections below).

The Advisory Board's recommendation in 1998 to allow the Director to embark on **fund raising** for research and development purposes has proven results with the attainment of the R9.07 million project from the Innovation Fund (Round 2) of DACST as reported in the previous annual report. As a direct consequence a number of other potential opportunities have opened up for further international investment into the field of Human Language Technologies (HLT) in South Africa to the order of approximately R70 - 140 million. (See 2.3.3.1)

2.2 Completed Projects

2.2.1 Modelling the intonation of imperatives in a Xhosa text-to-speech system

P Swart, J Louw, L Botha, & JC Roux

For a full description of this project - please see the Annual Report 1996/7

This project was completed in December 1999 by Mrs Louw (nee Ms Swart) and was awarded an MA (cum laude). Apart from setting specific guidelines of how to develop a TTS system in an African tone language, an important methodological point has been addressed regarding the nature of tagged speech when used as a basis for concatenation in speech synthesis.

Dissemination of results:

Previously reported:

i. Presentation at two day workshop at *9th International ALASA* conference, Durban (Swart, 1997)

ii. Presentation at *10th International Conference of the African Language Association of Southern Africa* (Swart, 1997)

New:

i. PH Swart. 2000. Prosodic features of Imperatives in Xhosa: Implications for a text-to-speech system. MA Thesis, University of Stellenbosch. 151 pp. Including CD-ROM with examples.

Forthcoming conference presentations:

i. October 2000**.** Synthesizing Prosody for Commands in a Xhosa TTS System. Philippa H. Louw, Justus. C. Roux & Elizabeth. C. Botha. Poster presentation at **ICSLP 2000: 6th International Conference of Spoken Language Processing,** Beijing, China.

ii. November 2000. Synthesizing Prosody for Commands in a Xhosa TTS System. Philippa H. Louw, Justus. C. Roux & Elizabeth. C. Botha. Paper to be read at **PRASA 2000: Eleventh Annual Symposium of the South African Pattern Recognition Association.** University of the Witwatersrand, Johannesburg. (Adapted version of (i))

This is the final report on this project.

2.3 Current projects

Projects are conducted within the following programmes:

- **Programme 1:** Linguistic projects

  (The phonetics-phonology interface)

- **Programme 2:** Resource development projects.

  (Resource development in African languages and capacity building)

- **Programme 3:** Language and speech technology projects

  (Human-machine interaction through African languages)

2.3.1 Programme 1: Linguistic projects

2.3.1.1 The production and perception of statement and question intonation in Xhosa

Ms Jackie Jones & JC Roux

For a full description of this project - please see the Annual Report 1996/7

This project is to be completed in **November 2000**. The external examination of this doctoral thesis will take place in November and it is expected that the candidate will graduate in December 2000. The project comprised detailed acoustic and perceptual analyses of the queclarative phenomenon in Xhosa focusing on linguistic, psycholinguistic and technological implications.

Dissemination of results:

Previously reported:

(i) One presentation at national conference of ALASA, Johannesburg: Jones (1996)

(ii) One presentation at regional workshop of ALASA, Pretoria: Jones (1996)

(iii) One presentation at an international workshop of ALASA, Durban (1997)

(iv) Two articles in an accredited journal; Jones, J, Louw, J & Roux, JC. (1998 and 1998a)

(v) One contribution to a Festschrift Jones, J & Roux, JC. (1998)

New:

Nothing at this stage. DLitt Thesis forthcoming.

2.3.1.2 The phonetics and phonology of click articulations in Nguni

JC Roux, G Dogil & M Jessen

For a full description of this project - please see the Annual Report 1996/7

This is an **ongoing study** with December 2001 as final date. Dr Michael Jessen, a member of the Institute for Computational Linguistics (IMS) at the University of Stuttgart joined in on this project as a postdoctoral researcher in 1998. The project has also drawn in a number of graduate students focusing on phonetic detail and phonological theories (cf the work of Naidoo mentioned below).

Dissemination of results:

Previously reported:

i. Presentations at five international conferences in, respectively, Vienna (Dogil, 1996)[1], Budapest (Roux, 1996), München (Dogil, 1996), Adelaide (Roux, 1996), Paris (Roux & Dogil, 1997) as well as at two national/regional conferences in South Africa: (Roux, 1996a, 1996b).

ii. Three publications as proceedings of international conferences: Dogil et al (1996), Dogil & Roux (1996), Dogil, Mayer & Roux (1998)

iii. One publication in an accredited journal: Roux & Lewis (1996)

iv. Two contributions to proceedings of international conferences: Roux & Dogil (1998), Dogil, Mayer & Roux (1998).

New:

Articles submitted for publication in accredited journals

- **Jessen, M & Roux, JC (*) Voice** characteristics associated with stops and clicks in Xhosa: An acoustic investigation. Submitted for publication to an international journal: *Journal of Phonetics*. (July 1999, previously listed). Accepted for publication pending smaller revisions (July, 2000)

- **Jessen, M (*)** An acoustic study of distinctive stop types and click accompaniments in Xhosa. Submitted for publication to an international journal: *Journal of Phonetics* (June, 1999), no feedback yet.

- **Naidoo, Shamila (*) Distinctive** feature theory: From linear to non-linear: An application to isiZulu. **Accepted** for publication in *South African Journal for African Languages* (SAJAL) in January 2001.

2.3.2 Programme 2: Resource development and capacity building projects

2.3.2.1 Preferred syllable structures in Nguni and Sotho

JC Roux, A Jones, J Louw, Dorothy Calata & Bongiwe Fingo

[1] Full details regarding these "older" references may be found in the 1997/8 Report.

For a full description of this project - please see the Annual Report 1996/7.

It was reported last year that this **ongoing study** was expected to be completed in **December 1999**, and that it would link on to the new DACST 2000 project. This has indeed been the case and the research is now as part of the African Speech Technology (AST) project.

The original Pattern Analysis Programme (PATANA) had to be refined for the AST project and now provides a better tool for this type of research.

Dissemination of results:

Previously reported

i. One presentation at an international conference: Vienna (Roux, 1996)

ii. Two presentations at national workshops (A Jones, 1997, 1997a)

In preparation:

Roux, JC & Louw, JAN. (*) Preferred syllable structures in Xhosa and their role in creating new lexical items. To be presented for publication in *South African Journal for African Languages*. (December 2000)

2.3.2.2 SASPEECH

From a practical point of view it has also become necessary to integrate this sub project into the AST project. A detailed description of the databases under development may be found in **Appendix A** (Progress Report 1 of the AST project, page 8-10). No further report will be forthcoming under this heading.

Dissemination:

Previously reported:

- **Roux, JC. 1998:** SASPEECH: Establishing speech resources for the indigenous languages of South Africa. *Proceedings of the First International Conference on Language Resources and Evaluation***.** (Eds A Rubio et al). Granada, Spain. European Language Resources Association - ELRA. Vol 1, 343-350.

2.3.2.3 Development of computer software

A wide range of new software tools is to be developed under the banner of the AST project. These tools will be specified in the appropriate Progress Reports of AST. These reports will obviously be made available to the Advisory Board.

A new Patana 2000 (Pattern Analysis) application was developed to to replace the older DOS-based Patana program for the transcription and statistical analysis of, *inter alia*, African languages. Patana 2000 is a Windows-based application developed by by Frank Olivier using Visual Basic 6.0. Using Patana 2000, a user can transcribe large ortographical texts in a short time using new transcription rules developed by Philippa Louw. The phonetic texts can then be analysed in a number of ways, using /CV/ (and more advanced) pattern searches with access to diphone and triphone lookups. The results can then be exported to other applications for further analysis.

2.3.2.4 Research capacity building projects

Previously reported

Three RCB projects with respectively the University of Venda, University of Transkei and the Qwa-Qwa branch of the University of the North were reported last year. Due to a lack of time no new projects of this nature have been embarked on. Specific capacity building related to the training of Mr J Masalesa (NRF, Internship) as well as to a number of graduate students involved in the projects are taking place on a continuous basis.

Result:

The RCB project at the HDIs resulted in a core of well skilled mother tongue speakers of, respectively, isiXhosa (Dr Jokweni, Mr Sotashe and Ms Qamata) and Sesotho (Mr Selebeleng), all who are actively involved in the new AST project.

2.3.3 Programme 3: Language and Speech Technology Projects

2.3.3.1 African Speech Technology Project (AST)

African Speech Technology (AST) is the working title of the DACST Innovation Fund Project (Round 2) awarded to RUEPUS in 1999, which officially commenced on January 1, 2000. The approved title of the project is *Promoting the development of the official languages of South Africa through language and speech technology applications*. This project is overseen by a Steering Committee comprising representatives of the consortium responsible for the project. It has a prescribed mode of report, notably the submission of quarterly progress reports, the first of which is included here as **Appendix B**. Over and above these mechanisms, DACST has appointed two Progress Monitors for each of their Innovation Fund projects. Monitors for the AST project are Prof M Kahn (Faculty of Education, UCT) and Prof E Blake (Department of Computer Science, UCT). A website (<http://www.ast.sun.ac.za/>) has been created for this project with links to and from the RUEPUS website.

As one of its deliverables, i.e. setting up national and international networks in order to stimulate an industry for Human Language Technologies in South Africa, the AST project embarked on forging contacts with a trust in Belgium in order to attract an investment of R70 million (with a further R70 million as partnership funds in SA) to this field. The negotiations are at a sensitive stage with the management of the Belgian Trust to visit South Africa (and the AST project) early in November 2000.

Dissemination:

Conference presentations:

i. Roux, JC & Louw, PH. 2000: Workshop and poster presentation entitled Black **South African English and Speech Technology Applications**. *International Conference on Linguistics in Southern Africa*. University of Cape Town.

ii. Roux, JC, Botha, EC, & Du Preez, JA. 2000. **Developing a multilingual telephone based information system in African languages**. *Second International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Athens, Greece.

iii. Schwardt L & Du Preez, J. 2000. **Mixed-order HMMs for language recognition**. *IEEE COMSIG-2000*, Stellenbosch.

Forthcoming conferences:

i. Philippa H. Louw, Justus. C. Roux & Elizabeth. C. Botha. (2000 October). **Synthesizing Prosody for Commands in a Xhosa TTS System.** Poster presentation at *ICSLP 2000: 6th International Conference of Spoken Language Processing,* Beijing, China

ii. Schwardt L & Du Preez, J. 2000 (October). **Efficient mixed-order hidden Markov model inference** *ICSLP 2000: 6th International Conference of Spoken Language Processing,* Beijing, China.

iii. Schwardt L & Du Preez, J. 2000 (October). **Automatic language identification using mixed-order HMMs and untranscribed corpora**. *ICSLP 2000: 6th International Conference of Spoken Language Processing,* Beijing, China.

Submission of article for publication in accredited journal:

i. Roux, JC & Louw, Phillipa H (2000) Speech **corpora for technology applications: Varieties of South African English**. Submitted for publication to *South African Journal of Linguistics.*

2.3.3.2 Language Engineering (LE): Principles and Practice

JC Roux, PH Louw, J Masalesa

A detailed description of this project is found in the 1997/8 Report.

The CSD/NRF awarded an Internship to RUEPUS to fund the position of Mr Masalesa who has spent the first year of internship working on this project. The project is making good progress in creating, *inter alia*, an awareness in the country towards the need for LE development. The material gathered in this project serves to assist in language planning with respect to the development of Human Language Technologies (HLT).

Two specific events directly and indirectly related to this project may be mentioned:

i. **Contribution to the formulation of language policy in South Africa**

The Director was invited to host a workshop at the Second National Language Indaba of the Department of Arts, Culture, Science and Technology (DACST) in Durban in May 2000 regarding the role of Language and Speech Technology in language policy in South Africa. He also presented a plenary lecture on this topic at the same conference. From both these events a number of specific recommendations followed, most of which eventually found their way into the *Final Draft Language Policy and Plan for South Africa* (5 June 2000) of the Advisory Panel on Language Policy which was presented to the Minister of Arts, Culture, Science and Technology and which is to be tabled in Parliament during this year.

ii. **Contribution to the formulation of a national strategy for Human Language Technologies in South Africa**

The Director was invited to join a national Steering Committee on Language and Information Technology set up by DACST and the Pan South African Language Board (PANSALB). He was eventually responsible for the drafting of a strategic planning document entitled *The Development of Human Language Technologies in South Africa: Strategic Planning*, which is currently under consideration by DACST and PANSALB. This strategic plan, if accepted, has far reaching implications for the development of National Lexicography Units (NLUs) and for the future of RUEPUS as well. Some more details on this are to be found in Section 4 below.

Dissemination:

National conference participation:

i. Roux, JC 2000 **Second National Language Indaba**. Organised by DACST and PANSALB, Holiday Inn, Durban

International conference participation

i. Masalesa, J & Roux, JC 2000. **Language Engineering in Southern Africa: a challenge for the new millennium.** *Interim Conference of the African Language Association of Southern Africa* **Gaborone**, Botswana.

2.4 Publications and conference presentations

2.4.1 Papers submitted to accredited journals

- **Jessen, M & Roux, JC (*) Voice** characteristics associated with stops and clicks in Xhosa: An acoustic investigation. (June, 1999) Accepted for publication in Journal *of Phonetics* pending final alterations.
- **Jessen, M (*)** An acoustic study of distinctive stop types and click accompaniments in Xhosa. Submitted for publication to an international journal: *Journal of Phonetics* (June, 1999), no feedback yet.
- **Naidoo, Shamila (*) Distinctive** feature theory: From linear to non-linear: An application to isiZulu. Accepted for publication in *South African Journal for African Languages* (SAJAL) in January 2001.
- **Roux, JC & Louw, Phillipa H** (*) Speech corpora for technology applications: Varieties of South African English. Submitted for publication to *South African Journal of Linguistics*

2.4.2 Publications in national and international conference proceedings

- **Roux, JC & Lewis, PW 1999.** On Xhosa L2 speech and intelligibility: Ejectives, implosives and clicks**.** *Proceedings of the 14th International Congress of Phonetic Sciences***,** San Francisco, USA; Regents of the University of California. CD-ROM format.
- **Roux, JC, Botha, EC, & Du Preez, JA, 2000**. Developing a multilingual telephone based information system in African languages. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens. Vol 2. pp 975-980. Also in CD-ROM format

2.4.3 Papers read at international conferences

2.4.3.1 Previously reported (overlap)

- **JC Roux. 1999.** *10th Biennial International Conference of the African Language Association of Southern Africa (ALASA 99).* UNISA, Pretoria.
- **Organiser and presentation of workshop:** *African Language Engineering in the New Millennium.*
- **Keynote speaker** on invitation*: Language and Technology: Challenges for African languages in the next millennium.*
- **JC Roux 1999**. Presentation at the *14th International Congress of Phonetic Sciences*, San Francisco, USA. Entitled **On Xhosa L2 speech and intelligibility: Ejectives, implosives clicks.**
- **Swart, PH**. 1999. Presentation at *10th International Conference of the African Language Association of Southern Africa* entitled **Prosodic features of Imperatives in Xhosa: A perceptual experiment.**

2.4.3.2 New

- **Roux, JC & Louw, PH.** 2000: Workshop and poster presentation entitled Black **South African English and Speech Technology Applications**. *International Conference on Linguistics in Southern Africa*. University of Cape Town.
- **Roux, JC, Botha, EC, & Du Preez, JA**. 2000. **Developing a multilingual telephone based information system in African languages**. *Second International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Athens, Greece.
- **Schwardt L & Du Preez, J**. 2000. **Mixed-order HMMs for language recognition**. *IEEE COMSIG-2000*, Stellenbosch.
- **Masalesa, J & Roux, JC** 2000. Language Engineering in Southern Africa: a challenge for the new millennium. *Interim Conference of the African Language Association of Southern Africa* **Gaborone**, Botswana

2.4.4 Papers read at national conferences

2.4.4.1 Previously reported (overlap)

- **Roux, JC. 1999 Introducing Translation Technology to the South African scene.** Invited plenary address at **National Seminar of the South African Translators` Institute.** ESKOM. Midrand.

2.4.4.2 New

- **Roux, JC 2000** Human Language technologies and its development in South Africa. Workshop at **Second National Language Indaba**. Organised by DACST and PANSALB, Holiday Inn, Durban

- **Roux, JC 2000** Language and Technology: Challenges for African languages in the new millennium. Plenary paper at **Second National Language Indaba**. Organised by DACST and PANSALB, Holiday Inn, Durban

2.5 Staff activities

2.5.1 Visits abroad

The Director made a visit in May 2000 to Athens, Greece in order to participate in the 2nd *International Conference on Language Resources and Evaluation*. A number of good contacts were made and it was during this conference that the Director was approached to serve in the International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (COCOSDA).

As part of the AST project the Director made a visit in June 2000 to Ieper in Belgium to visit Flanders Language Valley that was set up by Lernout and Hauspie. The aim of this visit was twofold: (i) to assess the nature of the technology available for possible implementation in the AST project and (ii) to negotiate with the S.AI.L[2]. Trust for possible investment in South Africa in the field of HLT. The negotiations are well underway and an announcement in this regard may be made before the end of this year. Should the negotiations be successful, it would open new research and development opportunities for, inter alia, linguists working in African languages.

2.5.2 Visitors

RUEPUS received visit from the following persons, for longer or shorter stays respectively. Staff members could interact with these visitors and the academic network of RUEPUS was strengthened considerably:

1. Prof. Ekkehard Wolff, Institut fuer Afrikanistik, University of Leipzig, Germany.
2. Prof. Arvi Hurskainen, Institute for Asian and African Studies, University of Helsinki, Finland.
3. Prof. Wim van Dommelen, National Technical University of Trondheim, Norway.
4. Prof. Katherine Demuth, Dept. of Cognitive & Linguistic Sciences, Brown University, Providence, USA.
5. Prof. Mark Johnson, Dept. of Cognitive & Linguistic Sciences, Brown University, Providence, USA.
6. Prof James Duplessis Emejulu, Omar Bongo University, Gabon. This visit was intended to arrange for an exchange programme for postgraduate students and researchers to the phonetics laboratory of RUEPUS.
7. Delegation from CIPAL (Belgium)…

Our paper is entitled and Steven Canon. Paper presented at the 31st Annual

African Linguistics Conference, Boston University, March 4, 2000).

[2] S.AI.L. is an acronym for **S**peech, **A**rtificial **I**ntelligence and **L**anguage.

2.6 Conferences and symposia presented

2.6.1 Previously reported

- Two day workshop on Language Engineering: *10th International Conference of the African Language Association (ALASA),* University of South Africa, Pretoria, July 1999
- Colloquium on experimental speech research, at the University of the North, Qwa-Qwa branch, Phuthaditjhaba, June 1999.

2.6.2 New

- Seminar on: Tagging and Glossing Sesotho, Mark Johnson and Katherine Demuth. Stellenbosch.

2.7 Implementation of research

Up till now research results have been disseminated as described above. Real world implementation of these results will take place with the development of an interactive multilingual information system.

3 INFRASTRUCTURE

3.1 Staff

The following persons were appointed for the periods specified:

Mr Jan Louw (M Eng.)	Ad hoc (1998) then 1/10/1998 - 30/11/1998
Ms Pippa Swart (BA Hons)	1/4/1998 - 31/3/1999
Ms D Calata	Ad hoc 1998/9
Ms B Fingo	Ad hoc 1998
Ms M Mdemka (BA)	Ad hoc 1998-9
Ms S du Plessis	1/6/1998 - 31/3/1999
Ms E van der Vyver (BA Hons)	1/6/1998 - 31/3/1999
Mr J Masalesa (BA Hons)	1/1/1999 - 31/3/1999 (31/12/1999)

RUEPUS was awarded an Internship from the CSD that made provision for the position of Mr Masalesa. He will be more fully employed in RUEPUS following this internship.

3.2 Resources

Resources for the study of both technology and phonetics are readily available in the different libraries on campus. RUEPUS furthermore has adequate access to the Internet.

3.3 Equipment

RUEPUS had to upgrade some of its computer equipment as routine practice. This is reflected in the financial report. At this point in time the laboratory is adequately equipped for running its own projects. Any new project sponsored from an external source will have to include equipment in its budget - this is indeed the case with the DACST 2000 project

where provision has been made for additional hardware and software.

4 PLANNING FOR NEXT YEAR (2000/2001 - YEAR 3 OF CYCLE 2)

The strategic plan accepted in 1998 will have to be adapted in view of the DACST 2000 project:

CYCLE 2	ACTIVITIES	STAFF	COSTS
Year 1 **1998/1999**	• Research in linguistic processing and technology development • Research capacity building projects with HDIs • Setting up networks at national and international levels • Initial fund raising initiatives	1 researcher (ling) 1 researcher (techn) Half day secretary	
Year 2 **1999/2000**	• Research in linguistic processing and technology development • Research capacity building projects with HDIs • Networking at national and international levels • Special awareness campaigns with major fund raising initiatives • Solicitation of projects • DACST (1998) proposal on speech recognition • SASPEECH	1 researcher (ling) 1 researcher (techn) Half day secretary	
Year 3 **2000/2001**	• ALEC partially in place (with RUEPUS projects) • **DACST 2000: (1/1/2000 - 31/12/2001)** • Research in linguistic processing and technology development • LE training / capacity building • **RUEPUS**: o **Programme 1:** Phonetics/phonology interface o **Programme 2:** Resource and software development Speech and text corpora o **Programme 3:** LE principles and practice	**Own Staff:** 30 Co-workers (nationally) 5 Co-workers (internationally) 1 Researcher (Linguistics) 1 Programmer Ad hoc subjects / co-workers	**For 2000**: R2.9 million **For 2001**: R2.9 million As per budget
Year 4 **2001/2002** (Final year for NRF funding)	• ALEC in place with final RUEPUS projects o Change of name at end of cycle in March 2002 o Host to DACST 2000 till end of 2002. • **DACST 2000: (1/1/2002 - 31/12/2002)** • Research in linguistic processing and technology development • LE training / capacity building • **RUEPUS / (S)ALEC:**[3] o **Programme 1:** Resource and software development Speech and text corpora o **Programme 2:** LE principles and practice	**Own Staff:** 30 Co-workers (nationally) 5 Co-workers (internationally) 1 Researcher (Linguistics) 1 Programmer (ad hoc) Ad hoc subjects / co-workers	**For 2001**: R3.1 million As per budget

The planning for the third year (2000/2001) below is in accordance with the proposed strategic plan as presented above:

4.1 Programme 1: Linguistic projects: The phonetics - phonology interface

4.1.1 The phonetics and phonology of click articulations in Nguni (Continuation)

JC Roux, G Dogil, M Jessen and M Jokweni

This is an ongoing study with estimated date of completion in 2001. Dr Michael Jessen, a member of the Institute for Computational Linguistics (IMS) and Dr Mbulelo Jokweni of UNITRA joined the team. The scope of the project has been expanded to include *inter*

[3] It is envisaged that (S)ALEC will continue after termination of NFR-funds at that point in time, but with independent funds originating from industry.

alia the study of ejectives and implosives in more detail. The relationship between phonetics and phonology will be explored with reference to these sound types as well as the implications of these sound types for automatic speech recognition and speech synthesis.

This project may only require new software - this is reflected in the budget.

4.2 Programme 2: Resource development projects

4.2.1 Preferred syllable structures in Nguni and Sotho (Continuation) - to be expanded to "SALText"

JC Roux, J Louw, James Masalesa & F Olivier

This project will be adapted to include so called "information mining" strategies in order to analyse the morphophonotactic structures of these languages in a systematic way. This will necessitate innovative software design to create computer-based tools for use by linguists. Mr Frank Olivier (a computer science undergraduate) has already proven his computational skills and will be responsible for these programmes.

The new acronym "SALText" refers to "South African Languages Text" corpora. Although the focus will exclusively remain on the Sotho and Nguni languages, it is envisaged that this project will eventually be expanded to all South African languages and that multilingual re-useable text resources and computer based access tools will be available for linguistic and technological applications.

Provision is made in the budget for

i. Language informants (subjects)

ii. Database labellers

iii. Computer programmer.

4.2.2 SASPEECH (Continuation)

JC Roux, J Louw, P Swart, M Jokweni, K Qamata, A Sotashe

This project will be continued with co-researchers from Unitra focussing on Xhosa.

Provision is made in the budget for language informants (subjects)

4.2.3 Text Corpora: Parsing Tools for the African Languages (On hold)

JC Roux, Marianna Visser, Kathy Demuth (Brown University, USA) and Annie Zaenen (Xerox, France)

This project that was announced in 1998 is still on hold depending on high-level negotiations with Xerox (France).

4.3 Programme 3: Language and Speech Technology Projects

4.3.1 Language Engineering: Principles and Practice (Continuation)

JC Roux, P Swart & J Masalesa

This project entails the continuous study of developments regarding LE technologies as well as of models of language and speech processing, and principles and policies underlying implementation.

No financial implications for the budget except for the position of Mr Masalesa.

4.3.2 DACST 2000 (Appendix A)[4]

Project co-ordinator: JC Roux

This project entitled *Promoting the development of the official languages of South Africa through language and speech technology applications*, is the result of a consortium proposal to the **Innovation Fund** of the Department of Arts, Culture, Science and Technology and amounts up to a total of R9 077 620,00 spread over a period of three years.

This can be seen as a major breakthrough towards the sensitising of Government for the role it has to play in developing the languages of the nation at technological level. A total of 35 co-workers will be involved, thirty of which will be South African citizens of which at least 15 will be mother tongue speakers of an African language. The activities will take place at various campuses across the country with RUEPUS as base.

This project is of further importance to RUEPUS in that it actually provides the financial backing for the transformation process of RUEPUS to ALEC as envisaged in the Strategic Plan accepted by the Advisory Board.

Detail may be found in Appendix A. The description below will be used by DACST as a press release when officially announcing the new successful projects:

University of Stellenbosch

Research Unit for Experimental Phonology -

Department of African Languages

Promoting the development of the official languages of South Africa through language and speech technology applications.

Proposal number: 21213

Focal area: Information Society

[4] Please see footnote 1 and the request for confidentiality.

Project duration: 3 years

Total funding: R 9,077,620

Funding year 1: R 2,983,750

Funding year 2: R 2,903,590

Funding year 3: R 3,190,280

This project aims to introduce the languages spoken in South Africa into the domain of interactive human-machine communication as the world moves into a technologically driven next millennium. In order not to become marginalized in any way it is necessary that the official languages spoken in this country be developed at technological level to meet the challenges of the information society.

It has become clear within developed countries that interactive voice enabled information systems functioning in particular languages are of the most powerful tools to impart information and knowledge on a broad basis within and across multilingual societies. These systems, which invoke **voice recognition, linguistic processing** and **speech generation** fall within the domain of Language and Speech Technology (LST) and find applications in all spheres of life. The single most important value of voice based LST applications is the fact that these systems are accessible to all people regardless their state of literacy as communication takes place in natural spoken language.

Apart from the implementation of an awareness campaign directed towards potential high profile end users of these systems, a **multilingual query and booking system** will be developed. This fully automated system will accept telephonic queries submitted in four languages, i.e. in either **South African English, isiXhosa, Sesotho** or **Afrikaans** after which it will access an information database and eventually supply the caller with the necessary information in good quality synthetic speech. Although the domain of the queries is initially restricted to **the hotel industry**, provision will be made for later "plug-in" modules, e.g. information on travel schedules, cinema and theatre programmes, insurance products, banking rates, automobile products, weather conditions etc.

The project will be conducted by a consortium of three interdisciplinary teams from respectively the universities of Stellenbosch (African Languages and Electronic Engineering) Pretoria (Electronic Engineering), Transkei (African Languages), and West Technology Holdings with co-workers from the universities of Potchefstroom, Port Elizabeth and the University of the North (Qwa-Qwa branch).

Project Co-ordinator: Prof. Justus Roux; **Phone**: +27 21 808 2017

jcr@maties.sun.ac.za

4.3.3 National and international conferences

As in the past provision is made in the budget for two national and one international conference attendance by staff members of RUEPUS in order to disseminate research results and to be involved in capacity building programmes.

4.4 Staff

The services of a young **linguistic researcher** is required to assist in theoretical work related to the linguistic project in Programme 1. It is important to keep abreast with developments in phonological theory and its language specific applications and in this regard the services of an MA student in African linguistics is urgently needed. Provision is made in the budget for such a person (Item 6). Two mother tongue speakers of African languages have indicated that they may be available for this position.

The Unit must have access to the services of a skilled computer scientist to develop software programmes for projects on Programmes 2 and 3. Mr Frank Olivier, a final year student in Computer Science have been involved in programme development within RUEPUS and the Department of African Languages during 1999. In order to retain his services for ad hoc programming tasks in 2000 (when he will be enrolled for an Honours degree), provision is made accordingly in the budget (Item 7)

Mr James Masalesa is a current Intern in the Unit with great research potential. In order to retain his services, provision is made for an amount of R50 000 in the budget (Item 8), which is on par with the amount he currently receives from the NRF.

Provision is also made for ad hoc co-workers / labellers / language assistants who have up to now fulfilled an extremely important task in the gathering of authentic speech material and/or in the perceptual testing of this data in experiments. These are normally undergraduate African students or Matriculation scholars from Khayamandi. It is envisaged that data gathering will need to be increased dramatically next year, therefore the anticipated need for 200 hours` participation (Item 9).

5 BUDGET

5.1 Statement of income and expenditure for the previous financial year (1998/9)

Please see **APPENDIX B**

5.2 Budget for next financial year (2000/2001)

Please see **APPENDIX C**

6 GENERAL

Thank you very much to the benefactors of RUEPUS, i.e. the NRF (CSD) and the Research Committee of the University of Stellenbosch, for continuous support over the years. It has enabled RUEPUS to grow in many ways to the point of recognition by Government in entrusting us with this DACST 2000 project from the Innovation Fund. Without your support,

as well as that of the Advisory Board and the dedicated work of all staff members this would not have been possible.

PROF. JC ROUX **STELLENBOSCH**

DIRECTOR: RUEPUS **29 OCTOBER 1999**