SOUTHERN AFRICAN LINGUISTICS AND APPLIED LANGUAGE STUDIES

Volume 21 2003

LANGUAGE TECHNOLOGY IN SOUTHERN AFRICA: RESOURCES AND APPLICATIONS

ISSN 1607–3614
EISSN 1727–9461

The *Southern African Linguistics and Applied Language Studies* is printed on an acid free, woodfree product, containing bagasse, which is the fibrous residue from processed sugar cane. The producers of this environment friendly paper, Sappi, have been certified as meeting the ISO 14001 environmental management system standard.

Layout, design, technical editing and production: NISC Pty Ltd, South Africa

Cover design: Goat Multimedia, Port Alfred, South Africa

Printing: Express Litho Services cc, Port Elizabeth, South Africa

AN INTRODUCTION TO 'HUMAN LANGUAGE TECHNOLOGY IN SOUTHERN AFRICA: RESOURCES AND APPLICATIONS

Gerhard B van Huyssteen
School of Languages, Potchefstroom University for CHE, Potchefstroom 2531, South Africa
e-mail: afngbvh@puk.ac.za

In February 2003, we sent out a call for contributions on 'Human Language Technology in South Africa: Resources and Applications', for a special issue of this journal. The invitation called for scholarly articles, reporting or reflecting on:

- the current state of the art of language technology resources and applications in South Africa;
- completed and ongoing research in Natural Language Processing, Corpus Linguistics, and Computational Linguistics, specifically with reference to the languages used in South Africa;
- the past, present, and future of language technology in South Africa.

In reply to this invitation we received sixteen contributions from scholars (and their local or international co-workers) working at most of the academic institutions with an interest in language technology in South Africa, resulting in a rather representative snapshot of local computational linguistic research.[1] We also received a contribution from two internationally acclaimed scholars (with no affiliation to any South African university), Eric Atwell and Bayan Shawar, on the development of a chatbot for Afrikaans.

After an intensive review process by local and international reviewers, ten contributions were eventually accepted for publication (excluding one project overview, and two reviews of books relevant to the computational linguistic community). The articles not only focus on all the official languages of South Africa, but also touch on most aspects of the research and development process of human language technologies. Consider Figure 1, which is a depiction of the research, development, and commercialisation process of human language technologies.

The first three steps in the developmental process have to do with resources. First, we have to determine standards, specifications, and protocols for the compilation of spoken and written corpora and other resources, with a view on the enabling technologies and applications to be developed eventually. After this has been done, one can start collecting the required corpora and compile relevant grammars, after which one engages in the process of developing enabling technologies (e.g. tokenisers, Named Entity Recognisers, Part of Speech taggers, etc.). Only then can one move on to the next important step in the developmental process, viz. the development of end-user applications. Of course, not all steps are always necessarily present in a certain project, and not essentially in this sequence, but at least it gives an overview of a typical developmental process.

The first 6 or 7 contributions in this issue deal with resources, of which the first three mainly concern the formulation of standards, specifications, and protocols for corpus compilation, as well as the eventual annotation, management, and exploration of corpora. Allwood and Hendrikse give a view on the design and process of, and requirements for compiling spoken language corpora for the nine official languages of South Africa, as well as the potential uses and applications of such corpora (one of only two contributions in this issue explicitly dealing with spoken language). Both the contributions by Van Rooy and Pretorius, and Allwood, Grönqvist and Hendrikse deal with the composition of word-class tagsets for the African languages of South Africa, with special reference to Setswana and isiXhosa respectively. All of these contributions, together with other articles in this issue, repeatedly stress either explicitly or implicitly the importance of and need for setting standards and protocols for corpus practice in South Africa — an issue that should be addressed sooner rather than later by the computational linguistic community and relevant authorities in South Africa.

Quite a few of the contributions in this issue deal with the development of enabling technologies

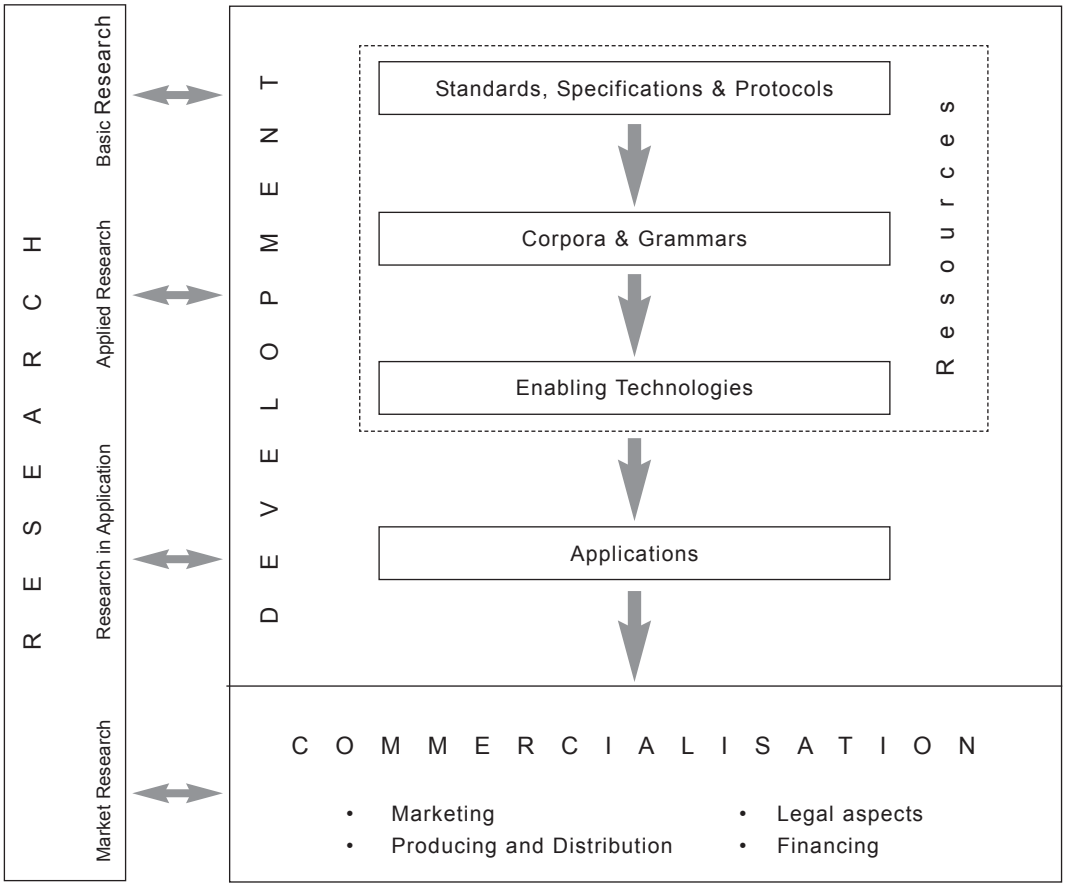

Figure 1: Research, Development, and Commercialisation of Human Language Technologies

(i.e. those natural language processing modules that can be used in the development of different applications). The abovementioned article by Allwood, Grönqvist and Hendrikse focuses on the development of Part-of-Speech taggers for the African languages (specifically for isiXhosa), while Joffe, De Schryver and Prinsloo discuss some computational issues that they came across while developing a software application for dictionary compilation, called TshwaneLex (which may be seen as either a kind of enabling technology, or an end-user application). Also within the scope of enabling technologies, Wissing reports on the development of a grapheme-to-phoneme converter for Afrikaans within the African Speech Technology project (the second contribution focusing explicitly on speech technology). In both the articles by Pretorius and Bosch, and Van Zaanen and Van Huyssteen, attention is focused on the development of lexica and morphological analysis tools for isiZulu and Afrikaans respectively. Pretorius and Bosch specifically points out the importance of such resources for the development of most natural language processing applications. The article by Shawar and Atwell also stresses the need for well-annotated (spoken) corpora for the development of enabling technologies and applications, in their case to develop a chatbot (i.e. a question-answer agent) for Afrikaans.

The last three contributions deal with the development and evaluation of end-user applications. Snyman and Naudé set forward a framework for the evaluation of a machine translation system under development at the University of the Free State, and subsequently report some preliminary results. Relating to this, the contributions by Prinsloo and De Schryver and Van Zaanen and Van

Huyssteen propose frameworks for the evaluation of spelling checkers for the South African languages. Additionally, Prinsloo and De Schryver give a description of the development and characteristics of their so-called first generation spelling checkers for ten of the South African languages, while Van Zaanen and Van Huyssteen report on their experiences in developing a so-called second generation spelling checker for Afrikaans.

To sum up, I think that it is clear from the above that many exiting things are happening in South Africa with regard to the research and development of human language technology resources and applications. Despite the fact that the fields of Human Language Technology, Computational Linguistics, Corpus Linguistics, and Natural Language Processing are still in their infancy in South Africa, much potential for further and ongoing work is expressed by the contributions in this special edition. I sincerely hope that this effort will inspire other scholars (locally and internationally) to become more interested in these fields of inquiry, specifically with regard to the South African context and the impact that Human Language Technology will have on our multilingual community. Maybe the day draws nigh to institute a society for Computational Linguistics in South Africa, in order to stimulate further work and incite more and more collaborative efforts. Then, with growing public and governmental interest and support, I only see a bright and shiny future for Human Language Technology in South Africa.

Finally, allow me a short word of appreciation to the following people:
- All contributors, for their hard work and enthusiasm;
- All reviewers, for their meticulous and comprehensive comments and suggestions;
- The editor-in-chief, Bertus van Rooy, for his encouragement, advice, and general hands-on help;
- The team at NISC, in particular Georgina Jones, for her patience and professional support; and
- Christo Els, for administrative assistance.

Notes

[1] Due to time constraints, colleagues from the University of Stellenbosch (who work on, among other projects, the seminal African Speech Technology project), were unfortunately unable to submit any contributions. Although the University of Stellenbosch is therefore not duly represented in this special edition, one of their co-workers from the Potchefstroom University for CHE, Daan Wissing, does report on aspects of the above-mentioned African Speech Technology project.