Voorwoord

Hierdie spesiale uitgawe van *Literator* word gewy aan mensetaaltegnologieë (MTT) vir Suid-Afrikaanse tale en bevat elf herbewerkte kongresreferate wat tydens spesiale temasessies by twee kongresse in 2007 in Suid-Afrika gelewer is. Hierdie twee kongresse was die gesamentlike kongres van die Linguistevereniging van Suider-Afrika (LVSA), die South African Applied Linguistics Association (SAALA) en die Suid-Afrikaanse Vereniging vir Taalonderrig (SAVTO), gehou op die Potchefstroomkampus van die Noordwes-Universiteit, en die 14de Internasionale Kongres van die African Language Association of Southern Africa (ALASA), gehou by die Nelson Mandela Metropolitaanse Universiteit, Port-Elizabeth.

Elke bydrae is onderwerp aan intensiewe, dubbel-anonieme ewekniekeuring deur drie lede van die keuringspaneel, bestaande uit nasionale en internasionale vakgenote.

Die artikels in hierdie publikasie verteenwoordig ’n breë spektrum van huidige navorsings- en ontwikkelingsaktiwiteite in die MTT-veld (spesifiek tekstegnologie) aan verskeie instellings in Suid-Afrika. Kwessies wat aangeraak word wissel van morfologiese analise tot masjienvertaling en MTT-projekbestuur. Die goue draad wat egter al die bydraes aan mekaar verbind, is die ontwikkeling van taalhulpbronne vir hulpbronarm tale, herbruikbaarheid, die deel van hulpbronne tussen verwante tale, asook standaardisering met die oog op volhoubare ontwikkeling.

Die eerste vyf artikels in hierdie spesiale uitgawe handel oor aspekte van morfologiese analise van drie Suid-Afrikaanse tale. Twee bydraes (wat albei oor Sothotale handel) skakel in by ’n projek oor die rekenaarmatige morfologiese analise van verskeie Afrikatale. In hulle artikel, *Towards a computational morphological analysis of Setswana compounds*, ondersoek Pretorius, Viljoen, Pretorius en Berg die vorming van naamwoord + naamwoord-samestellings met behulp van rekenaarmatige middele ten einde te verstaan hoe hierdie proses geformaliseer, gemodelleer en in ’n morfologiese analiseerder vir Setswana geïmplementeer moet word. Anderson en Kotzé

fokus in hulle bydrae, *Verbal extension sequencing: an examination from a computational perspective*, op die voorkoming van oormatige generering in die morfologiese analise van veelvoudige reekse van werkwoorduitgange in Sesotho sa Leboa. Maniere om uitgangkombinasies te beperk, word ondersoek deur toetsdata uit literatuursoektogte, leksikografie-ondersoeke en rekenaarmatige morfologiese analises van tekste te gebruik.

Drie verdere artikels handel oor aspekte van outomatiese Afrikaanse morfologiese analise. Die bydrae van Pilon, Puttkammer en Van Huyssteen getiteld, *Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans*, jukstaponeer twee verskillende benaderings tot die ontwikkeling van kerntegnologieë vir hulpbronarm tale, te wete 'n reëlgebaseerde benadering en 'n masjienleerbenadering. Hulle toon aan dat die gebruik van geheuegebaseerde masjienleer suksesvol gebruik kan word om spesifieke noodsaaklike hulpbronne vir Afrikaans te ontwikkel.

Die ander twee artikels fokus op aspekte van fleksie in Afrikaans. Groenewald en Van Huyssteen in die artikel, *Outomatiese lemma-identifisering vir Afrikaans*, gee eers 'n bondige oorsig oor die belangrikste kwessies met betrekking tot fleksie in Afrikaans, waarna hulle die ontwikkeling van 'n lemma-identifiseerder vir Afrikaans bespreek. Deur ook gebruik te maak van geheuegebaseerde masjienleer, slaag hulle daarin om 'n lemma-identifiseerder wat 92,8% akkuraat is, te ontwikkel. In haar bydrae getiteld, *Die ontwikkeling van 'n fleksievormgenereerder vir Afrikaans*, gebruik Pilon eweneens geheuegebaseerde masjienleermetodes om 'n "omgekeerde lemma-identifieerder" (d.i. 'n kerntegnologie wat verskillende woordvorme van 'n gegewe lemma kan genereer) te ontwikkel. Sy rapporteer belowende resultate, met aanduidings hoe die akkuraatheid verder verbeter kan word.

Verwant aan kwessies van morfologiese analise is sake wat betrekking het op hulpbronne en tegnologieë vir morfosintaktiese annotering/etikettering. In hulle artikel, *On the development of a tagset for Northern Sotho with special reference to the issue of standardisation*, argumenteer Taljard, Faaß, Heid en Prinsloo dat veelvlakkige annotering noodsaaklik is om voorsiening te maak vir die morfologiese kompleksiteit van Sesotho sa Leboa. Hulle oorweeg ook verder aspekte van standaardisering teen die agtergrond van hergebruik, die deel van hulpbronne, en die moontlike aanpassing vir gebruik deur ander disjunktief-geskrewe Suid-Afrikaanse Bantoetale.

In hierdie spesiale uitgawe word vir die eerste keer verslag gedoen oor die ontwikkeling van woordnette vir Suid-Afrikaanse tale. Na ’n algemene agtergrond oor woordnette, doen Kotzé in sy artikel, *Ontwikkeling van ’n Afrikaanse woordnet: metodologie en integrasie*, ’n metodologie vir die semi-outomatiese konstruering van sinoniemstelle vir Afrikaans aan die hand. In die artikel, *Derivational relations in English, Czech and Zulu wordnets,* beskryf Bosch, Fellbaum en Pala inisiatiewe om morfologiese en semantiese reëlmatighede in Engels, Tsjeggies en isiZulu vas te vang ten einde dit in die rekenaarmatige verwerking en konstruering van woordnette te gebruik, veral vir isiZulu en die ander Bantoetale.

In die domein van taalidentifisering beskryf Zulu, Botha en Barnard in hulle bydrae getiteld, *Orthographic measures of language distances between the official South African languages*, twee verskillende metodes om die ooreenkomste en verskille tussen Suid-Afrikaanse tale te meet. Deur gebruik te maak van n-gramstatistiek en die Levenshtein-afstandsmeting is hulle in staat om tale effektief en objektief in families wat ooreenstem met algemene taalkundige kennis te groepeer.

Die ontwikkeling en evaluering van ’n eindgebruikertoepassing in die veld van masjienvertaling word beskryf in die artikel, *An overview of the EtsaTrans machine translation system: compilation of an administrative domain,* deur Ehlers en Hanekom. Hierdie masjienvertaalsisteem is tans die enigste een wat in Suid-Afrika ontwikkel word vir ’n gespesialiseerde domein waar dokumente herhaaldelik gebruik word, te wete dokumentasie vir vergaderings.

Die bydrae van Janke getiteld, *Besigheidsprosesbestuur in mensetaaltegnologiehulpbrononontwikkeling: ’n gevallestudie*, kan gesitueer word in die onontginde terrein van mensetaaltegnologiebestuur. Sy argumenteer dat metodes uit die praktyk van besigheidsprosesbestuur aangewend kan word in die bestuur van die ontwikkeling van mensetaaltegnologieë en demonstreer haar voorgestelde raamwerk aan die hand van ’n gevallestudie. Sy kom tot die gevolgtrekking dat die voorgestelde raamwerk kan bydra tot die standaardisering van volhoubare tegnologieë.

Hierdie spesiale uitgawe van *Literator* sluit af met ’n tweetalige terminologielys van die belangrikste terme in die rekenaarlinguistiek. Hierdie lys kan gesien word as ’n hulpmiddel in die vestiging van Rekenaarlinguistiek as ’n veeltalige dissipline in die Suid-Afrikaanse konteks.

Ten slotte wil ons graag ons dank uitspreek teenoor alle outeurs vir hulle waardevolle bydraes, lede van die ewekniekeurderspaneel vir hulle gedetailleerde kommentaar en konstruktiewe voorstelle, asook die *Literator*-redaksiespan vir hulle bystand en steun.

Gerhard B. van Huyssteen & Sonja E. Bosch
Gasredakteurs: spesiale uitgawe van *Literator*

Preface

This special issue of *Literator* is dedicated to human language technologies (HLT) for South African languages and contains eleven reworked conference papers that were read during special theme sessions at two conferences in South Africa during 2007. The two conferences were the joint conference of the Linguistics Society of Southern Africa (LSSA), South African Applied Linguistics Association (SAALA) and South African Association for Language Teaching (SAALT), held on the Potchefstroom Campus of the North-West University, and the 14th International Conference of the African Language Association of Southern Africa (ALASA), held at the Nelson Mandela Metropolitan University, Port Elizabeth.

Each paper submitted was subjected to intensive double-blinded peer reviewing by three members of a review panel consisting of national and international scholars.

The articles in this publication represent a broad spectrum of current research and development activities in the field of HLT (in particular text technologies) at various institutions in South Africa. Issues that are addressed range from morphological analysis to machine translation and HLT project management. However, the golden thread running through all these issues is development of language resources for resource-scarce languages, reusability, sharing of resources for related languages, and standardisation for purposes of sustainable development.

The first five articles in this special issue deal with aspects related to the morphological level of analysis of three South African languages. Two contributions both dealing with Sotho languages, are linked to a project on computational morphological analysis of several African languages. In their article *Towards a computational morphological analysis of Setswana compounds*, Pretorius, Viljoen, Pretorius and Berg investigate the formation of noun + noun compounds by computational means in order to understand how this process should be formalised, modelled and subsequently implemented in a morphological analyser for Setswana. Anderson and Kotzé, in their contribution, *Verbal extension sequencing: an examination from a com-*

putational perspective, focus on the prevention of over-generation in the morphological analysis of multiple verbal extension sequences in Sesotho sa Leboa. Means of limiting the extension combinations are investigated by using test data from literature research, lexicographic investigation and the computational morphological analysis of texts.

Three further articles cover aspects of automatic Afrikaans morphological analysis. The contribution of Pilon, Puttkammer and Van Huyssteen entitled, *Die ontwikkeling van 'n woordafbreker en kompositumanaliseerder vir Afrikaans*, juxtaposes two different approaches to the development of core technologies for resource-scarce languages, namely a rule-based approach and a machine learning approach. It is shown that specifically memory-based machine learning can be used successfully to develop essential resources for Afrikaans.

The other two articles focus on aspects of inflection in Afrikaans. In the article, *Outomatiese lemma-identifisering vir Afrikaans*, Groenewald and Van Huyssteen firstly give a concise overview of the most important issues regarding inflection in Afrikaans, after which they discuss the development of a lemmatiser for Afrikaans. By additionally using memory-based machine learning, they succeed in developing a lemmatiser with 92,8% accuracy. In her article, *Die ontwikkeling van 'n fleksievormgenereerder vir Afrikaans*, Pilon also uses memory-based machine learning methods to develop an "inverted lemmatiser" (i.e. a core technology that can generate various word forms from a given lemma). She reports promising results and indicates how accuracy can be improved.

Closely related to issues of morphological analysis, are resources and technologies dealing with morphosyntactic annotation/tagging. In their article, *On the development of a tagset for Northern Sotho with special reference to the issue of standardisation*, Taljard, Faaß, Heid and Prinsloo argue for multilevel annotation in order to account for the morphological complexity of Sesotho sa Leboa. They also consider aspects of standardisation against the background of reuse, sharing of resources, and possible adaptation for use by other disjunctively written South African Bantu languages.

This special issue also sees the first reporting on the development of wordnets for South African languages. After a background discussion on wordnets, Kotzé reports on a proposed methodology for the semi-automatic construction of synsets for Afrikaans in his article, *Ontwikkeling van 'n Afrikaanse woordnet: metodologie en integrasie.*

In their contribution, *Derivational relations in English, Czech and Zulu wordnets,* Bosch, Fellbaum and Pala describe efforts to capture the morphological and semantic regularities of derivational processes in English, Czech and Zulu for purposes of exploiting them for suitable computational processing and wordnet construction, especially with regard to Zulu and other Bantu languages.

Within the domain of language identification, Zulu, Botha and Barnard describe in their contribution, *Orthographic measures of language distances between the official South African languages*, two different methods for measuring the differences and similarities between the South African languages. Using n-gram statistics and the Levenshtein distance measure, they are able to effectively and objectively cluster languages in family groups corresponding to our general linguistic knowledge.

The development and evaluation of an end-user application in the field of machine translation is described in the contribution, *An overview of the EtsaTrans machine translation system: compilation of an administrative domain*, by Ehlers and Hanekom. This machine translation system is currently the only one being developed in South Africa for a specialised domain with documents that are repetitive in nature, viz. documentation of meetings.

The contribution of Janke entitled, *Besigheidsprosesbestuur in mensetaaltegnologiehulpbronontwikkeling: 'n gevallestudie*, falls within the unexploited field of HLT management. She argues that methods from the practice of business process management can be used in the management of HLT development and demonstrates her proposed framework by means of a case study. She concludes that the proposed framework could contribute to the standardisation of sustainable technologies.

This special issue of *Literator* concludes with a terminology list of the most important terms in computational linguistics. This list can be regarded as a contribution for the establishment of Computational Linguistics as a multilingual discipline in the South African context.

Finally we would like to express our gratitude to all authors for their valued contributions, to the members of the peer review panel for their detailed comments and constructive suggestions, as well as to the *Literator* editorial team for their assistance and co-operation.

Gerhard van Huyssteen & Sonja Bosch
Guest Editors: *Literator* special issue