Die ontwikkeling van ’n omvattende masjienleesbare leksikon vir Zoeloe

Author:
Ronelle van der Merwe[1]

Affiliation:
[1]School of Computing, University of South Africa, South Africa

Correspondence to:
Ronelle van der Merwe

Email:
vdmerwer@unisa.ac.za

Postal address:
PO Box 7426, Westgate 1734, South Africa

How to cite this abstract:
Van der Merwe, R., 2011, ‘Die ontwikkeling van ’n omvattende masjienleesbare leksikon vir Zoeloe’, *Suid-Afrikaanse Tydskrif vir Natuurwetenskap en Tegnologie* 30(1), Art. #270, 1 page. http://dx.doi.org/10.4102/satnt.v30i1.270

Note:
This abstract was initially presented as a paper at the annual Natural Sciences Student Symposium, presented under the protection of the *Suid-Afrikaanse Akademie vir Wetenskap en Kuns*. The symposium was held at the University of Pretoria on 05 November 2010.

The following members formed part of the committee that was responsible for arranging the symposium: Mr. R. Pretorius (Department of Geography, University of South-Africa), Dr E. Snyders (NECSA), Dr M. Landman (Department of Chemistry, University of Pretoria) and Dr W. Meyer (Department of Physics, University of Pretoria).

The development of a comprehensive machine-readable lexicon for Zulu
A comprehensive machine-readable (MR) lexicon is a strategic resource in the development of language technology. Two essential aspects are the development of a comprehensive data model and its accurate and efficient implementation. This study focuses on the development of MR lexicons for the South African Bantu languages and more specifically Zulu.

’n Omvattende masjienleesbare (ML) leksikon is ’n strategiese hulpbron in die ontwikkeling van taaltegnologie. Hierdie studie fokus op die ontwikkeling van ML-leksikons vir die Suid-Afrikaanse swart tale en in die besonder vir Zoeloe. Ten einde ’n ML-leksikon vir Zoeloe te ontwikkel benodig ons enersyds ’n omvattende datamodel en andersyds die leksikale inligting self. Die doel van hierdie studie is die ontwikkeling en implementering van die datamodel. Ideaal gesproke sal hierdie omvattende ML-leksikon die basiese digitale bergingsfasiliteit vir alle beskikbare leksikale inligting vir Zoeloe wees. Hierdie inligting behoort in ’n formaat beskikbaar te wees wat die hergebruik daarvan in ander taaltegnologieë en natuurliketaalverwerkingstoepassings, soos ’n morfologiese analiseerder, ondersteun.

Die agglutinerende aard en komplekse rekursiewe morfologiese strukture van Zoeloe bring mee dat ’n unieke datamodel ontwikkel moet word vir die omvattende Zoeloe ML-leksikon. Die betekenisse van woorde in Zoeloe verander deur ’n verskeidenheid van agter- en voorvoegsels by te voeg. So sal die woordwortel *bon* verskeie modifikasies kan ondergaan met telkens nuwe betekenisse. Verder is die volgorde van die byvoegsels ideosinkraties en die volgorde behoort uit reeds beskikbare data verkry en in die ML-leksikon gestoor te word.

Enkele voorbeelde van *bon* (‘sien’) met toelaatbare agtervoegsels sluit die volgende in:

1. *-bon-is-*: toon, laat sien
2. *-bon-an-*: sien mekaar
3. *-bon-is-an-*: toon aanmekaar
4. *isi-bon-is-el-o-ana*: klein voorbeeld

Hierdie agglutinering van morfeme kan beskou word as ’n vorm van rekursie en behoort in die datamodel, sowel as die implementering daarvan in die ML-leksikon, verteenwoordig te word.

Laastens word implementeringsmoontlikhede waarin rekursie ondersteun word, bespreek. Die keuse van die benadering of databasistipe word onder andere bepaal deur die verskillende *perspektiewe* op die data, die *toegangmeganismes* tot die data en die *tipe* data wat gestoor word. Weens die strategiese aard van die ML-leksikon as hulpbron is dit noodsaaklik dat dit verskeie perspektiewe en toegangsmeganismes toelaat ten einde die optimale inbedding in ’n verskeidenheid toepassings te ondersteun. Leksikale data is semi-gestruktureerd en die implementeringsbenadering sal vir rekursie voorsiening moet maak. XML en Unicode is *de facto*-standaarde vir annotering en dekodering. Ten opsigte van implementering is die fokus op die gebruik van suiwer XML (Native XML) en XML-versoenbare databasisse.

Hierdie studie vind plaas binne die konteks van die ISO FDIS 24613:2008 konsep-standaard vir leksikale annoteringsraamwerke. Die uitkomste van die studie is drieërlei: Eerstens die ontwikkeling van ’n omvattende datamodel vir die ML-leksikon, tweedens ’n ondersoek na die mees geskikte benaderings vir implementering van die ML-leksikon en derdens die bou en kritiese evaluering van ’n prototipe.