Ontwikkeling van ’n Afrikaanse woordnet: metodologie en integrasie

G. Kotzé
Sentrum vir Tekstegnologie (CTexT)
Potchefstroomkampus
Noordwes-Universiteit
POTCHEFSTROOM
E-pos: gidi8ster@gmail.com

Abstract

Development of an Afrikaans wordnet: methodology and integration

The Afrikaans wordnet is a lexical-conceptual network in the form of an electronic lexical database, developed at the North-West University. In this article, a methodology for a semi-automatic construction of the entries – so-called synonym sets – is investigated. Firstly, a background is given on the nature of a wordnet, as well as “WordNet”, on which it is based. Other wordnets, as well as applications of wordnets, are also discussed here. Next, the macrostructure of a wordnet in terms of its integration and compatibility with other wordnets is investigated, after which the proposed methodology is presented with a discussion of the results. Finally, a projection is made to the integration of the Afrikaans wordnet with other resources, which include “WordNet” and an Afrikaans lexical database, called ALEXANDER.

Opsomming

Ontwikkeling van ’n Afrikaanse woordnet: metodologie en integrasie

Die Afrikaanse woordnet is ’n leksikaal-konseptuele netwerk in die vorm van ’n elektroniese leksikale databasis wat aan die Noordwes-Universiteit ontwikkel is. In hierdie artikel word ’n metodologie vir ’n semi-outomatiese konstruksie van die inskrywings – sogenaamde sinoniemstelle – bekyk. Daar word eers ’n agtergrond gegee oor wat ’n woordnet is, sowel as oor “Word Net” waarop dit gebaseer is. Ander woordnette, sowel as toe-

passings van woordnette, word ook bespreek. Daar word voorts gekyk na die makrostruktuur van 'n woordnet in terme van sy integrasie en verenigbaarheid met ander woordnette, waarna die voorgestelde metodologie voorgelê en die resultate bespreek word. Laastens is daar 'n vooruitskouing van die integrasie van die Afrikaanse woordnet met ander bronne, waaronder "WordNet" en 'n Afrikaanse leksikale databasis genaamd ALEXANDER.

1. Inleiding

Die Afrikaanse woordnet is 'n leksikale databasis, befonds deur die Departement van Wetenskap en Tegnologie en ontwikkel aan die Noordwes-Universiteit. Hierdie artikel is gebaseer op 'n deel van die databasis wat voorlopig afgehandel is en wat bestaan uit 5 041 inskrywings. Figuur 1 vertoon 'n paar addisionele statistieke. Tydens die skrywe hiervan is die woordnet in die proses om uitgebrei te word na 10 068 inskrywings. In hierdie artikel word 'n metodologie vir 'n semi-outomatiese konstruksie van die inskrywings, sogenaamde sinoniemstelle, bekyk waarna 'n vooruitskouing volg op die integrasie van die Afrikaanse woordnet met ander bronne. In die volgende afdeling word eers besin oor wat 'n woordnet is, met verwysing na *WordNet* waarop dit gebaseer is, ander woordnette en enkele toepassings.

Figuur 1: Statistiek van die Afrikaanse woordnet (2006-2007)

	Selfstandige naamwoorde	**Byvoeglike naamwoorde**	**Werkwoorde**	**Totaal**
Sinoniemstelle	3 294	298	1 449	5 041
Woordeenhede (nie uniek gesorteer nie)	7 583	687	4 604	12 874
Semantiese verhoudings: absoluut[1]				11 561
Semantiese verhoudings: werklik[2]				6 017

1 Hierdie syfer word uit die XML (Extensible Markup Language) gelees. Sekere verhoudings, soos antonieme, is simmetries (werk na albei kante toe) en word daarom twee keer getel, terwyl andere, soos die hiponiem-/hiperoniemverhouding, slegs na een rigting werk en dus slegs een keer per verhoudingspaar getel word (in die XML word hiponieme geïmpliseer deur slegs na die hiperonieme te verwys).

2. *WordNet*/woordnet – wat is dit?

'n Woordnet is 'n leksikaal-konseptuele netwerk in die vorm van 'n elektroniese databasis. Die prototipe is *WordNet* (Fellbaum, 1998), wat in die vroeë negentigerjare aan die Princeton universiteit (VSA) as 'n psigolinguistiese model van die mentale leksikon ontwikkel is. *WordNet* is 'n nuttige verwysingsbron, en blyk ook nuttig vir tegnieke in natuurliketaalprosessering soos byvoorbeeld woordbetekenisvereenduidiging (Morato *et al*., 2004:273-274). Tydens die skryf van hierdie artikel bevat die nuutste weergawe van *WordNet*, *WordNet 3.0*, reeds 117 659 inskrywings.

2.1 Struktuur van *WordNet*

WordNet is deels gebaseer op teorieë uit die relasionele leksikale semantiek (vgl. Mel'čuk, 1996 en Evens, 1988), waar betekenis verteenwoordig word deur netwerke van nodusse wat bestaan uit woorde of woordgroepe wat aan mekaar verbind is. In *WordNet*, en by uitbreiding natuurlik die Afrikaanse woordnet, word só 'n nodus verteenwoordig deur 'n sogenaamde sinoniemstel (Engels: *synonym set*, of kortweg, *synset*). 'n Sinoniemstel is dus 'n groep sinonieme wat saam een konsep verteenwoordig. Hierdie sinoniemstelle word dan deur semantiese verhoudings verbind om een groot netwerk te vorm. Só kan die Afrikaanse sinoniemstel *motor, motorkar, kar* deur 'n hiponimiese verhouding met die sinoniemstel *voertuig, vervoermiddel* verbind word. Dit wil sê: 'n *motor* is 'n soort *voertuig*, waar *motor* die hiponiem en *voertuig* die hiperoniem is. Die meeste sinoniemstelle in *WordNet* het definisies, een of meer opsionele voorbeeldsinne en domeinetikette, soos byvoorbeeld *Sociology*. Die sinoniemstelle is gestruktureer in 'n ontologie, die *Suggested Upper Merged Ontology* (SUMO), wat iets van die semantiese aard van die sinoniemstel beskryf, byvoorbeeld *Body Motion* of *Transaction*. Pease *et al*. (2002) verduidelik in meer besonderhede oor hierdie *IEEE*-inisiatief[3] wat vir die semantiese web ontwikkel is. Uit bogenoemde is dit dus logies dat 'n woordnet slegs uit inhoudswoorde kan bestaan, met ander woorde selfstandige naamwoorde, werkwoorde, byvoeglike naamwoorde en bywoorde.

---

2 Hierdie verhoudings verwys na bestaande Afrikaanse sinoniemstelle. Omdat die *Princeton WordNet* se struktuur oorgeneem word, bevat die XML ook verwysings via ID-nommers na sinoniemstelle wat nog nie bestaan nie.

3 Die *Institute of Electrical and Electronics Engineers*, 'n niewinsgewende professionele organisasie vir die bevordering van tegnologie, volgens hulle amptelike webblad.

Verskeie ander semantiese verhoudings kom ook voor, soos byvoorbeeld *cause*, waar een aksie 'n ander logies meebring, soos wat die geval is met *snork* en *slaap*. Vir 'n meer volledige lys, kyk Fellbaum (1998:109).

2.2 Uitbreiding van *WordNet*

Toe *WordNet* gratis op die internet beskikbaar geword het, het dit in 'n kort tyd tot 'n globale fenomeen uitgegroei. Dit het betekenis op 'n unieke manier verteenwoordig en was enig in sy soort. Dit was veral die toepassingsmoontlikhede van *WordNet* in natuurliketaalprosessering wat wye belangstelling gelok het. In sulke gevalle word dit dikwels gekombineer met ander bronne soos korpora. Dit is byvoorbeeld reeds in ooreenstemming gebring met *Roget* se Tesourus (Mandala *et al*., 1999) en die bogenoemde *SUMO*-ontologie.

Kort voor lank het ander instansies ook woordnette vir hulle tale begin ontwikkel. Die eerste groot mylpaal was die ontwikkeling van *EuroWordNet* vanaf 1996-1999, 'n woordnetdatabasis vir agt Europese tale: Nederlands, Frans, Duits, Engels, Tsjeggies, Spaans, Italiaans en Estnies. Die verskillende woordnette is met mekaar belyn deur die gebruik van 'n sogenaamde *Intertalige indeks* (*Inter-Lingual-Index*, of kortweg *ILI*) wat bestaan uit 'n lys sinoniemstelle uit *WordNet*. Elke sinoniemstel in elke woordnet het 'n unieke identifikasienommer. Indien daardie nommer ooreenstem met 'n nommer wat aan 'n sinoniemstel in die indeks gekoppel is, dui dit konseptuele ekwivalensie aan. Deur die indeks te gebruik as skakel tussen die verskillende woordnette, is dit moontlik om spesifieke betekenisse van woorde of woordgroepe in die ander tale op te soek (Vossen, 1999).

BalkaNet (2001-2004) was nog 'n omvangryke multitalige woordnetprojek. Die databasis bevat woordnette vir Grieks, Turks, Serwies, Roemeens, Tsjeggies en Bulgaars (Tufiş *et al*., 2004:9). Die oogmerke van die projek sluit 'n nog groter dekking oor die verskillende tale heen in as met *EuroWordNet*; maksimale oorvleueling en verenigbaarheid met die reeds bestaande woordnette in *EuroWordNet* en die ontwikkeling van toepassings in die gebied van woordbetekenisvereenduidiging (*word sense disambiguation*); die intelligente indeksering van dokumente en intertalige inligtingsontsluiting (*cross-lingual information retrieval*) (Tufiş *et al*., 2004:11, 13-14).

2.3 Toepassings van woordnette

Behalwe as leksikale verwysingsbronne vir menslike gebruikers, word woordnette ook in natuurliketaalprosessering op verskeie gebiede toegepas. Een hiervan is sogenaamde konseptuele vereenduidiging of woordbetekenisvereenduidiging. Dit word deur Morato *et al*. (2004:273) gedefinieer as “precision and relevance in response to a query via resolution of semantic inconsistencies”. Dit word vir verskeie soorte toepassings aangewend, soos die vereenduidiging van internet-soektogvrae deur middel van programme soos *SimpliFind* en *Oingo* (Morato *et al*., 2004:273).

Inligtingsonttrekking is nog ’n toepassing wat ’n groot rol in die organisasie en voorstelling van inligting op die internet speel volgens Morato *et al*. (2004:272). Een gebruik daarvan is byvoorbeeld by vraaguitbreiding (*query expansion*) in internet-soekenjins wat nog ’n belangrike toepassing van ’n woordnet is. ’n Voorbeeld van so ’n enjin is *simpli.com* (http://www.simpli.com/), wat deur George Miller ontwikkel is en wat tegnologie bevat wat later deur *Google* (http:// www.google.com/) gebruik is in die vorm van ’n produk, *Google AdSense* (https://www.google.com/ adsense/). Buscaldi *et al*. (2004) toon hoe *WordNet* gebruik kan word om geografiese terme in soektogte uit te brei om meer gepaste resultate te lewer.

In die toepassing van outomatiese kategorisering en strukturering van dokumente word algoritmes gebruik om sekere sleutelwoorde en hulle relevansie te soek om dokumente te kan klassifiseer. Dit speel ’n belangrike rol in data-ontginning en daar bestaan verskeie produkte wat hierdie tegnologie aanwend, soos byvoorbeeld *PolyAnalyst* (*MegaPuter Intelligence*, http://www. megaputer.com/). Maniere waarop *WordNet* aangewend word in hierdie proses sluit die onttrekking van semantiese eienskappe deur middel van die grammatikale kategorisering van selfstandige naamwoorde, werkwoorde en byvoeglike naamwoorde in, asook die gebruik van die konseptuele verteenwoordiging van kennis in *WordNet* om modelle vir die voorspelling van die relevansie van die betrokke data te bou (Morato *et al*., 2004:274).

WordNet kan ook as ’n bron in masjienvertalingsprosesse gebruik word. Kim *et al*. (2002) het byvoorbeeld ’n algoritme voorgestel om *WordNet* en datagedrewe modelle te gebruik om teikenwoorde in masjienvertaling te kies. In ’n meer onlangse artikel (Lee *et al*., 2004:207-210) word aangetoon hoe *WordNet* gebruik word om Koreaanse meervoudige nominatiewe naamvalskonstruksies suksesvol in masjienvertaling te prosesseer.

As hulpmiddel vir die opsporing van beelde en klankmonsters op die internet kan *WordNet* ook 'n rol speel. *MultiMediaMiner* (Morato *et al*., 2004:275) is 'n voorbeeld van 'n prototipesisteem wat hierdie taak verrig.

WordNet word ook saam met korpora gebruik om die juiste betekenis van woorde vas te stel en dit is ook al belyn met ander leksikale bronne (byvoorbeeld *Roget's Thesaurus*). Korpora word dikwels ook semanties geannoteer, waar elke inhoudswoord aan 'n sinoniemstel in die woordnet gekoppel is (vgl. byvoorbeeld Fellbaum *et al*., 2001).

In die volgende afdeling word die makrostruktuur van 'n woordnet in terme van verenigbaarheid en integrasie met ander woordnette bespreek.

3. Makrostruktuur

'n Woordnet se makrostruktuur kan enige aantal en versameling sinoniemstelle bevat. In *EuroWordNet* en *BalkaNet* is daar op 'n lys konsepte besluit wat heel eerste in al die betrokke woordnette ingesluit moet word. Hierdie konsepte kom in ten minste twee van die betrokke tale voor en het meer semantiese verhoudings met mekaar en ander sinoniemstelle as die konsepte wat nie hier ingesluit word nie. Die redes vir die insluiting van hierdie konsepte is dat dit belangrik is om goeie interlinguistiese verenigbaarheid en integrasie te verseker, en dat dit 'n goeie kernwoordnet skep met baie verbindings wat reeds met vrug gebruik kan word. Dit lê 'n goeie grondslag vir uitbreiding.

Bogenoemde konsepte word basiskonsepte (*base concepts*) genoem (Global Wordnet Association, 2008). Dié wat in die *BalkaNet*-projek gebruik is, is gratis op die internet beskikbaar, en daarom is daar besluit om die eerste kernmakrostruktuur hierop te baseer. Baie ander woordnette het ook hierdie benadering of 'n variasie hierop gevolg, soos byvoorbeeld die Nederlandse woordnet. In die Nederlandse woordnet is die basiskonsepte nie net op bogenoemde maatstawwe gekies nie, maar ook op hulle hiërargiese posisie in 'n reeds bestaande Nederlandse leksikale databasis en op grond van hoe dikwels hulle daarin voorkom (Vossen *et al*., 1999:22-23).

Dit is dus duidelik dat 'n tipiese woordnet veral aan die begin 'n klein algemene standaardleksikon bevat. Dit kan natuurlik uitgebrei word tot so ver as wat die taal se leksikon dit toelaat. Die huidige *Word-*

Net 3.0 bevat byvoorbeeld 117 659 sinoniemstelle – verreweg die meeste van alle woordnette.[4]

In die volgende afdeling word ’n paar outomatiese metodes bespreek wat in die bou van ’n woordnet toegepas kan word.

4. Outomatiese metodes

WordNet is ’n voortdurende projek waar die sinoniemstelle, ten minste aan die begin, met die hand gebou is. Vanaf die begin van die konstruksie van *EuroWordNet*, is daar egter verskeie outomatiese metodes ontwikkel om ten minste ’n gedeelte van ’n woordnet outomaties te kan skep. Dit word die meeste toegepas in sogenaamde uitbreidingsmetodologieë, wat die vertaling van sinoniemstelle vanaf ’n brontaal na die taal wat die nuwe woordnet verteenwoordig, behels. ’n Samesmeltingsmetodologie het die teikentaal as beginpunt: die sinoniemstelle word eers in die eie taal geskep en dan later aan ’n bronwoordnet gekoppel. ’n Uitbreidingsmetodologie blyk volgens Piek Vossen, projekkoördineerder van *EuroWordNet* (1996-1999), minder kompleks te wees en kan die hoogste graad van verenigbaarheid oor verskillende woordnette heen verseker. Vir die bou van die Afrikaanse woordnet is dus besluit op ’n uitbreidingsmetodologie, waar die sinoniemstelle vanaf Engels vertaal word, aangesien Afrikaans en Engels konseptueel naby aan mekaar is: albei is Wes-Germaanse tale, Engels het baie invloed op Afrikaans gehad en hulle sprekers deel ook baie aspekte van hulle kulture as gevolg van die gemengde Britse en Nederlandse koloniale geskiedenis van Suid-Afrika. Afrikaans het ook ’n hoogs ontwikkelde woordeskat met ekwivalente vir baie Engelse gespesialiseerde terme, soos wat afgelei kan word uit die bestaan van woordeboeke soos die *Pharos Afrikaans-Engels/English-Afrikaans-woordeboek/dictionary* (Du Plessis, 2005) (meer as 200 000 trefwoorde)[5] en die *Verklarende Woordeboek van die Afrikaanse Taal* (Van Schalkwyk, 2005) (nagenoeg 186 000 trefwoorde).[6]

Die *Princeton WordNet* (nuutste weergawe is 3.0:2006) is die enigste woordnet wat gratis beskikbaar is en ook die volledigste, daarom

4 Die syfer word gegee deur die amptelike webblad, http://wordnet.princeton/edu/ onder *WordNet statistics*.

5 Volgens die amptelike webblad van Pharos.

6 Volgens die inleiding op die CD-ROM (2005).

is dit gekies as bronwoordnet. Vir die voorgestelde metodologie word die elektroniese weergawe van Eksteen (1997) se *Groot Woordeboek* gebruik. 'n Proeflopie is met die hand gedoen, aangesien die regte vir die outomatiese gebruik van die woordeboek eers onlangs verkry is. Die fokus in hierdie artikel is op die resultate van hierdie proeflopie, wat as 'n klein eerste weergawe dien van 'n "goudstandaard" wat in die toekoms gebruik sou kon word om outomatiese metodes verder te toets.

Verskeie goeie uitbreidingsmetodologieë bestaan om 'n woordnet te bou. Sommige van die metodes in hierdie metodologieë is egter nie altyd moontlik om toe te pas nie, omdat elke metodologie in 'n unieke omgewing met unieke bronne geskep is. Die sogenaamde *gloss matching*-metode in Pianta *et al*. se *MultiWordNet*-metodologie vereis byvoorbeeld 'n komplekse outomatiese verwerking van woordeboekglosse (Pianta *et al*., 2002:296), wat tydrowend is en dus nie altyd moontlik toepasbaar is nie as gevolg van tydsbeperkings. Ander metodes is weer taamlik algemeen toepasbaar en vereis slegs 'n hoëkwaliteit tweetalige woordeboek. Daar is besluit om 'n versameling metodes wat aan laasgenoemde vereiste voldoen, te kombineer en hulle effektiwiteit teen 'n goudstandaard te toets. In hierdie geval is die goudstandaard 'n stel sinoniemstelle wat met die hand vertaal is en wat uit selfstandige naamwoorde bestaan. Die metodes word op dieselfde stel toegepas en met die goudstandaard vergelyk om agter te kom hoe akkuraat die outomatiese vertalings is. Die resultate word in afdeling 5 voorgelê.

Metodes uit twee verskillende metodologieë word hier getoets, naamlik dit wat in *MultiWordNet,* 'n multitalige woordnetdatabasis (Pianta *et al*., 2002) gebruik is en een van die Roemeense woordnette (Barbu & Mititelu, 2005). Die metodes word vervolgens bespreek asof dit 'n outomatiese proses is, al het dit vir hierdie ondersoek handmatig geskied. In die volgende onderafdeling word die *MultiWordNet*-metodes bespreek.

4.1 Metodes uit *MultiWordNet*

Hierdie metodes, soos beskryf in Pianta *et al*. (2002), is slegs getoets op selfstandige naamwoorde, maar kan in beginsel op enige woordsoort toegepas word, aangesien dit slegs staatmaak op die vertalings van woorde in 'n tweetalige woordeboek. Die toepassings van die metodes lewer as resultaat 'n lys van Engelse sinoniemstelle. Langs elkeen van hierdie sinoniemstelle is 'n stel Afrikaanse woorde en hulle betekenisnommers, wat as vertaalekwivalente voorgestel word. Langs elke Afrikaanse woord is ook 'n vertrouenstelling

(*confidence score*) in persentasievorm. Hierdie telling is ’n poging om die waarskynlikheid van ’n woord om aan die naasliggende Engelse sinoniemstel gekoppel te wees, te voorspel en is gebaseer op die metodes wat voorts beskryf gaan word. Elke lys van Afrikaanse woorde (per Engelse sinoniemstel) word, ideaal gesproke, gesorteer van hoog tot laag op die vertrouenstelling om dit vir die leksikograaf wat die Afrikaanse sinoniemstelle moet bou, makliker te maak om te besluit watter woorde in die Afrikaanse sinoniemstel ingesluit moet word. Indien die vertrouenstellings akkuraat genoeg is, kan hierdie metode gebruik word en sal dit die woordnet se konstruksie aansienlik versnel. Die akkuraatheid van hierdie metode, soos toegepas op Afrikaanse data, word in afdeling 5 bespreek. (Kyk Figuur 1 aan die einde van afdeling 4.1 vir ’n voorbeeld van ’n Engelse sinoniemstel met voorgestelde Afrikaanse woorde en hulle vertrouenstellings.)

As voorbereiding vir die toepassing van hierdie metodes is ’n lys gemaak van die woorde in die Engelse sinoniemstelle wat in al bogenoemde metodes toegepas gaan word. Vir hierdie proeflopie bestaan dit uit ’n stel van 40 sinoniemstelle wat almal selfstandige naamwoorde bevat. ’n Tweede lys is ook gemaak wat bestaan uit al die Afrikaanse vertalings van hierdie woorde. Die volgende twee metodes word toegepas:

- **sinoniemstelsnypunt** (*synset intersection*): Vir elkeen van bogenoemde Afrikaanse woorde, is die volgende gedoen:

  - ’n Lys van Engelse vertalings van die woord is gegenereer met behulp van die *Groot Woordeboek*. By die Afrikaanse woord *skoot* word byvoorbeeld die volgende vertalings gegenereer: *shot*, *report*; *blast*; *time*, *turn*; *fold*; *sheet* (*ship*); *lap*; *womb*; *bosom*.

  - Een van die betekenisgroepe van hierdie vertalings is geneem. ’n Betekenisgroep is soos ’n sinoniemstel omdat dit ’n lys vertalings is, maar net van een betekenis van die woord. Sinonieme binne betekenisgroepe word deur kommas geskei, maar betekenisgroepe self deur kommapunte, soos byvoorbeeld met *shot*, *report*; *blast* (as vertalings van *skoot*) waar *shot* en *report* sinonieme is, maar *blast* nie ’n sinoniem is van *shot* of *report* nie. Die eerste betekenisgroep van die lys vertalings van *skoot* is dus *shot*, *report*.

  - Vervolgens is ’n lys gegenereer van al die Engelse sinoniemstelle waarin *shot* of *report* voorkom. Hoe meer woorde in ’n

sinoniemstel ooreenstem by bogenoemde betekenisgroep, hoe groter is die vertrouenstelling (*confidence score*) vir die Afrikaanse woord *skoot* om gekoppel te word aan die genoemde sinoniemstel. Byvoorbeeld pas een woord in die betekenisgroep *shot*, *report*, naamlik *shot*, by een van die woorde in die Engelse sinoniemstel *shooting*:1, en *shot*:3 – dus is die ooreenkoms 1 uit 2 woorde. Dit is dus logies om af te lei dat daar 'n redelike kans is vir die woord *skoot* om deel te vorm van 'n Afrikaanse sinoniemstel wat konseptueel ekwivalent is aan *shooting*, *shot*. Soos meer woorde deur hierdie proses gaan, het *shooting*, *shot* uiteindelik 'n hele lys van sulke woorde en die woorde met die beste vertrouenstellings behoort dié te wees wat in die Afrikaanse sinoniemstel kom. Die volgende formule word vir die berekening van die vertrouenstelling voorgestel:[7]

$CS1 = \frac{99i}{l}$, waar l die aantal woorde of woordgroepe in die Engelse sinoniemstel is en i die aantal woorde of woordgroepe in die betekenisgroep met ekwivalente in die sinoniemstel. Indien die twee getalle volledig ooreenstem, is die telling op sy hoogste, naamlik 99%. Daar is besluit om 'n persentasie naby aan, maar nie gelyk aan 100% nie te kies, bloot om aan te dui dat dit die resultaat is van 'n outomatiese proses en dat dit op hierdie stadium nie menslik gekontroleer is nie. Die resultaat van bogenoemde voorbeeld is dus 99 * 1/2 = 49,5%.

- Die volgende metode is *truvertaling* genoem na aanleiding van die Engelse *back translation*. Dit word apart van die eerste metode toegepas en die resultate van albei metodes word gekombineer om 'n finale vertrouenstelling te verkry. Truvertaling word soos volg toegepas: Uit die eerste lys Afrikaanse woorde, wat bestaan uit al die vertalings van die stel Engelse sinoniemstelle, word die volgende vir elke Afrikaanse woord gedoen:

  - Een van die betekenisgroepe van al die vertalings van die woord word geneem. As voorbeeld word die woord *gedaante* geneem. Die lys van vertalings is die volgende: *shape*, *form*, *aspect*; *configuration*; *spectre*, *apparition*; *face*.

7 Pianta *et al.* (2002) gee geen formules in hulle artikel nie. Hierdie formule is die outeur se eie voorstel.

- Die eerste betekenisgroep van die vertalings: *shape*, *form*, *aspect* word weer geneem.

- Weereens word ’n lys van al die Engelse sinoniemstelle met die woorde *shape*, *form* of *aspect* in gegenereer.

- Vir elke sinoniemstel word ’n vertrouenstelling uitgewerk op grond van die aantal woorde of woordgroepe in die sinoniemstel wat na *gedaante* terugvertaal kan word. Die sinoniemstel *shape*, *form* se woorde kan byvoorbeeld albei na *gedaante* terugvertaal word.

- Die formule wat hier voorgestel word, is so te sê dieselfde as bogenoemde: $CS2 = \frac{99t}{l}$, waar *l* die aantal woorde of woordgroepe in die Engelse sinoniemstel is en *t* die aantal woorde of woordgroepe in die sinoniemstel wat na die Afrikaanse woord terugvertaal kan word. In die geval van *shape, form* kan albei woorde terugvertaal word en is die vertrouenstelling dus 99*2/2 = 99%.

Die gemiddelde van bogenoemde vertrouenstellings word bereken om ’n finale vertrouenstelling uit te werk, dus:
$CombinedCS = \frac{CS1 + CS2}{2}$ waar *CS1* die sinoniemstelsnypuntmetode en *CS2* die truvertaalmetode se vertrouenstellings onderskeidelik is. Die finale resultaat bestaan, soos reeds aan die begin van hierdie afdeling genoem, uit Engelse sinoniemstelle met een of meer Afrikaanse woorde en betekenisnommers waaraan vertrouenstellings toegeken is. Hier volg ’n voorbeeld van toegekende vertrouenstellings vir die Engelse sinoniemstel *shape*:2, *form*:6

Figuur 2: Vertrouenstellings vir Afrikaanse woordbetekenisse aan ’n Engelse sinoniemstel toegeken

Engelse sinoniemstel	Afrikaanse woorde	Vertrouenstelling
shape:2, form:6	gedaante	99,00
	vorm(1)	74,25
	vorm(2)	74,25
	fatsoen	74,25
	gestalte	24,75

4.2 Metodes uit 'n Roemeense woordnet

Ter ondersteuning van die resultate word daar nog metodes uit 'n ander metodologie op dieselfde sinoniemstelle toegepas. Dit kom uit een van die Roemeense woordnette.[8] Daar word voorts op die volgende twee metodes gefokus, wat deur Barbu en Mititelu (2005) heuristiese reëls genoem word. Die toepassing van elke reël skep reeds 'n sinoniemstel in plaas van om vir elke woord 'n vertrouenstelling toe te ken en is dus volledig outomaties. Die resultate word vergelyk met 'n goudstandaard en die metodes word daarvolgens geëvalueer. Vir die steekproef word presies dieselfde data as vir die vorige stel metodes gebruik. In die volgende onderafdeling volg 'n beskrywing van die Roemeense metodes.

4.2.1 Roemeense metodologie: heuristiese reël 1

Vir elke Engelse sinoniemstel, word die volgende gedoen:

- Indien ten minste een woord of woordgroep in die sinoniemstel eenduidig is, met ander woorde slegs een betekenis het, word 'n Afrikaanse sinoniemstel geskep wat bestaan uit al die vertaalekwivalente van die betrokke woord of woordgroep. Die eenduidigheid van 'n woord of woordgroep word vasgestel deur te kyk na die aantal sinoniemstelle wat die betrokke woord of woordgroep bevat. Indien daar slegs een is, word hierdie reël toegepas. 'n Voorbeeld hiervan is die sinoniemstel *living thing:1, animate thing:1*, waar albei woordgroepe slegs in hierdie sinoniemstel voorkom.
- Indien al die woorde in die Engelse sinoniemstel polisemies of meerduidig is, bevat die Afrikaanse sinoniemstel daardie woorde wat in al die sinoniemstelwoorde se vertalings voorkom. As voorbeeld dien die Engelse sinoniemstel *discovery*:1, *find*:2, *uncovering*: 2. Die lys van vertalings is:
  - van discovery: ontdekking, vonds; openbaring; ontknoping; ooplegging (van dokumente)
  - van find: vonds, ontdekking; vangs

8 Dit is die tweede Roemeense woordnet, waar gepoog is om die eerste een wat in die *BalkaNet*-projek gebou is, as goudstandaard te gebruik om bestaande metodes te verbeter (kyk Barbu & Mititelu, 2005 vir meer inligting).

- *uncovering* is nie ’n trefwoord in die *Groot Woordeboek* nie. Dit word dus nie in ag geneem nie.

- Twee woorde, naamlik *ontdekking* en *vonds*, kom in albei groepe voor, dus vorm hulle die sinoniemstel wat, volgens die reël, gelyk is aan *discovery*:1, *find*:2, *uncovering*:2.

4.2.2 Roemeense metodologie: heuristiese reël 2

Die tweede reël maak voorsiening vir die feit dat ’n konsep in een taal slegs meer of minder spesifiek in die ander taal uitgedruk kan word. ’n Voorbeeld is die Engelse sinoniemstelle *election*:2 (met die betekenis van om iets te kies) en *choice*:2, *selection*:1, *option*:3, *pick*:9 wat albei deur die Afrikaanse woorde *keuse*, *kiesing* en *verkiesing*, wat in dieselfde sinoniemstel sal staan, vertaal word. Hierdie reël word soos volg toegepas:

- Vir elke Engelse sinoniemstel, word die volgende gedoen:
  - Soos in die eerste reël, word al die vertalings van al die woorde of woordgroepe gegenereer in die Engelse sinoniemstel. Vir *choice*:2, *selection*:1, *option*:3, *pick*:9 word byvoorbeeld die volgende vertalings gegenereer:
    - *choice*: *keus(e)*; *keur*; *verkiesing*; *beste*, *fynste*
    - *selection*: *keuse*, *seleksie*; *uitkiesing*; *keuring*; *keur(spel)* (musiek); *versameling*
    - *option*: *opsie*, *keuse*, *voorkeur*; *verkoopreg*; *verkiesing*; *reg van keuse*
    - *pick*: *keuse*; *beste*
  - Die Afrikaanse sinoniemstel word soos in die eerste reël geskep, naamlik die versameling van al die woorde wat in al die vertalinggroepe voorkom. In bogenoemde geval is die enigste woord wat aan hierdie vereiste voldoen *keuse*.
  - Presies dieselfde word met ’n hiperoniem of ’n hiponiem van die Engelse sinoniemstel gedoen, soos dit in *WordNet* voorkom. As voorbeeld dien ’n hiponiem van *choice*:2, *selection*:1, *option*:3, *pick*:9, naamlik *election*:2. Omdat hierdie sinoniemstel slegs uit een woord bestaan, bevat die snypuntstel outomaties al die vertalings van die stel, naamlik *keuse*; *kiesing*; *verkiesing*, *eleksie*.

- Die resultaat is twee Afrikaanse sinoniemstelle. Indien daar woorde in die twee sinoniemstelle is wat in albei voorkom, word daar nou 'n nuwe Afrikaanse sinoniemstel geskep wat uit hierdie woorde bestaan. In die twee bogenoemde sinoniemstelle is daar 'n woord wat in albei voorkom, naamlik *keuse*, en dus word 'n nuwe sinoniemstel geskep wat bestaan uit *keuse*. Die sinoniemstel word nou aan die twee Engelse sinoniemstelle gekoppel, naamlik die huidige een en die hiperoniem of hiponiem.

Die sinoniemstel *election*:2 bestaan uit slegs een woord en dus word hierdie metode in werklikheid nie toegepas nie, aangesien dit volgens die outeurs slegs geldig is vir sinoniemstelle wat uit twee of meer woorde of woordgroepe bestaan. Daar was egter in die steekproef nie 'n sinoniemstel waar die reël toegepas kon word sonder om ook 'n sinoniemstel met slegs een woord of woordgroep te betrek nie. Hierdie was dus slegs 'n voorbeeld om die reël te demonstreer.

Barbu en Mititelu voorsien 'n komplikasie in bogenoemde metode. As daar byvoorbeeld 'n sinoniemstel *A* in die bronwoordnet (Engels) bestaan wat 'n hiperoniem is van sinoniemstel *B*, wat weer 'n hiperoniem is van sinoniemstel *C*, en sinoniemstelle *A* en *B* is gekoppel aan sinoniemstel *S* in die teikenwoordnet (Afrikaans), terwyl sinoniemstelle *B* en *C* gekoppel is aan sinoniemstel *T* in die teikenwoordnet, word die volgende gedoen: Die sinoniemstel in die teikenwoordnet met die diepste vlak in die hiërargie word gekies, met ander woorde *T*, en dit word aan al drie sinoniemstelle in die bronwoordnet gekoppel. Bogenoemde situasie is volgens die outeurs ongewens en die reël vir die toewysing is soos volg: "... choose the assignment that maximizes the sum of depth level of the two synsets" (Barbu & Mititelu, 2005:102). In die Afrikaanse steekproefdata was dit nie nodig om hierdie reël toe te pas nie.

Wanneer bogenoemde Roemeense metodes klaar toegepas is, word die resultate apart gehou. Die geskepte sinoniemstelle word vergelyk met dié van die goudstandaard – in hierdie voorbeeld die genoemde 40 sinoniemstelle – en word geëtiketteer as reg of verkeerd op grond van die ooreenkomste tussen die twee stelle. Daar is vyf verskillende scenario's, wat Barbu en Mititelu (2005) soos volg beskryf:

- Die sinoniemstelle (geskepte vs. goudstandaard) is gelyk (*identical*).

- Die geskepte sinoniemstel se woorde vorm ’n deelversameling van dié van die goudstandaard (*under-generation*).
- Die geskepte sinoniemstel het al die woorde of woordgroepe van die goudstandaard en nog meer (*over-generation*).
- Die geskepte sinoniemstel en die goudstandaard deel sekere woorde of woordgroepe en ander nie (*overlap*).
- Die geskepte sinoniemstel het geen woorde of woordgroepe wat in die goudstandaard is nie (*disjoint*).

Slegs die eerste twee scenario’s, *identical* en *under-generation*, word as korrek beskou.

5. Bespreking van resultate

Dit blyk uit die resultate dat die Roemeense metodes nie baie geslaagd was in hierdie beperkte eksperiment nie (kyk Figuur 2). Vergeleke met die goudstandaard, is slegs sewe sinoniemstelle uit die 40 reg, waarvan ses as *under-generation* geëtiketteer is. Sewe uit die 40 is egter nie in berekening gebring nie, omdat die sinoniemstel uit slegs een woord of woordgroep bestaan. Die reëls kan volgens die outeurs slegs op sinoniemstelle toegepas word wat twee of meer woorde of woordgroepe bevat. Dit beteken dus dat sewe uit 33 reg is. Die persentasie korrekte sinoniemstelle is egter nog steeds slegs 21,21%.

Figuur 3: Opsomming van die resultate van die Roemeense metode

Aantal sinoniemstelle	***Identical***	***Under-generation***	***Over-generation***	***Overlap***	***Disjoint***	**Nie van toepassing**
40	1	6	0	1	25	7
	2,50%	15,00%	0,00%	2,50%	62,50%	17,50%
		17,50%			65,00%	
	Reg		**Verkeerd**			

Reg: sonder die laaste kolom (“nie van toepassing”) – 21,21% (dus uit 33)

By die *MultiWordNet*-metodes se resultate is daar ook inkonsekwenthede. Hierdie metodes word geëvalueer deur die Engelse sinoniemstelle se Afrikaanse ekwivalente in die goudstandaard te vergelyk met die lys met vertrouenstellings wat vir die betrokke Engelse sinoniemstel gegenereer is. Woorde uit die goudstandaard

behoort hoë vertrouenstellings te kry, maar dit is ongelukkig dikwels nie die geval nie. Die gemiddelde vertrouenstelling met woorde uit die goudstandaard is 46,5%. Die volgende probleme was faktore in al bogenoemde metodes:

- Engelse sinoniemstelle wat uit slegs een woord bestaan, kan makliker verkeerd vertaal word, aangesien daar nie ander woorde is om die konsep ondubbelsinnig te maak nie. Soms is dit egter die teenoorgestelde: die vertrouenstelling is 99% al behoort die woord nie in die sinoniemstel nie. Daar is byvoorbeeld meer as een sinoniemstel *walk*, elk met 'n ander betekenis. Die een met die betekenis *wandeling* kan dus moontlik verkeerdelik vertaal word met *wandelpad*, wat nog 'n vertaalekwivalent van *walk* is.
- Engelse sinoniemstelle bevat soms woordgroepe waarvoor daar nie in 'n standaard tweetalige woordeboek ekwivalente is nie. 'n Voorbeeld is *living thing*:1, *animate thing*:1 wat dus noodwendig 'n vertrouenstelling van 0 moet kry, en as *disjoint* geëtiketteer moet word.
- Die woordeboek gee soms nie genoeg vertaalekwivalente nie, of bevat eenvoudig nie die woorde wat vertaal moet word nie. Die woorde *interaksie* en *kognisie* kom byvoorbeeld glad nie voor nie, asook in Engels die woorde *buss* en *out-migration*.
- Werkwoorde wat as selfstandige naamwoorde gebruik word, soos *voetslaan* en *skiet*, word dikwels nie as sodanig deur 'n woordeboek weergegee nie, omdat so 'n woordsoortverandering 'n reëlmatige proses in Afrikaans is en afgelei kan word. Dit word uitgesluit om plek te bespaar.
- Die woordeboek is nie altyd konsekwent met terugvertalings nie. Die Engelse woord *shape* vertaal byvoorbeeld onder andere na *gestalte*, maar *gestalte* gee nie *shape* as 'n vertaalekwivalent aan nie.
- Meer ingewikkelde metodes wat uit dieselfde metodologieë kom, is nie hier toegepas nie. Dit behels byvoorbeeld die outomatiese ontleding van definisies en so meer.

Nog 'n probleem is dat daar nie altyd Afrikaanse vertalings is nie, soos byvoorbeeld vir die bofbalterm *base on balls*. Die Engelse sinoniemstel moet dus in sulke gevalle gemerk word as *nie-gelek-sikaliseerd*. Dit is trouens 'n algemene probleem en kan deels toegeskryf word aan die feit dat *WordNet* 'n databasis van Amerikaanse Engels is, terwyl plaaslike tweetalige woordeboeke soos die *Groot*

Woordeboek eerder fokus op Brits-Suid-Afrikaanse Engels soos dit oor die algemeen in Suid-Afrika gebruik word. Alhoewel hierdie twee variante in hulle standaardvorms grootliks ooreenstem, is daar na aanleiding van die outeur se ervaring met handmatige vertalings, genoeg verskille om ’n outomatiese metodologie negatief te affekteer.

Vertrouenstellings is dus voorlopig, volgens die resultate, nie goeie aanduidings van die geskiktheid van woorde om in sinoniemstelle opgeneem te word nie, omdat goeie woorde soms lae tellings kry. Iets wat nog nie getoets is nie, is of ongeskikte woorde ook hoë tellings kan kry. Indien dit nie die geval is nie, kan die tellings moontlik steeds gebruik word, omdat hoë tellings dan ’n goeie aanduiding sal wees van geskikte woorde, behalwe as die sinoniemstel uit slegs een woord of woordgroep bestaan. Sulke eenwoordsinoniemstelle moet voorlopig met die hand gekorrigeer word.

Die steekproeflys is verdeel tussen 20 sinoniemstelle wat heeltemal lukraak gekies is, en 20 wat bestaan uit drie of meer woorde wat geen woordgroepe of frases soos *living thing* bevat nie. Die gemiddelde vertrouenstelling vir die eerste lys (54,32) is heelwat hoër as vir die tweede lys (35,25), maar dit kan wees omdat daar heelwat eenwoordsinoniemstelle in die eerste lys is wat natuurlik hoë tellings gekry het, omdat die woorde almal uit die goudstandaard kom. In die eerste lys is slegs een sinoniemstel wat volgens die eerste Roemeense reël reg is, terwyl daar ses regte gevalle in die tweede lys is. Die verskil is miskien nie so beduidend nie, aangesien frases in die eerste lys wat geen woordeboekvertalings het nie as verkeerd gemerk is, terwyl hulle natuurlik nie in die tweede lys voorkom nie.

Omdat die Roemeense metodes ook swak gevaar het, moet hulle resultate vergelyk word met die ooreenstemmende vertrouenstellings. Hier is egter ook inkonsekwenthede: Sommige sinoniemstelle wat “reg” is, het lae vertrouenstellings, en sommiges wat “verkeerd” is, het weer hoë tellings. Die gemiddelde vertrouenstelling van “verkeerde” sinoniemstelle is egter heelwat laer (32,55) as dié wat “reg” is (45,25), alhoewel albei tellings laag is. Op hierdie stadium is dit nie duidelik wat ’n goeie telling is nie en verdere toetse moet eers gedoen word. Sodra dit vasgestel is, kan ’n beduidende deel van die woordnet moontlik outomaties of ten minste semi-outomaties geskep word. Die toepassing van verskillende metodes kan gekombineer word deur die resultaat te neem van die metode wat die beste teen ’n (meer volledige) goudstandaard gevaar het. Dit kan veral van nut wees indien ander metodes gebruik word. ’n Voorbeeld is metodes uit die uitbreidingsmetodologie van die Spaanse

woordnet in *EuroWordNet* (Atserias *et al*., 1997), wat moontlik in die toekoms toegepas sou kon word. Bogenoemde beste resultaat kan dan ter ondersteuning met *MultiWordNet*-vertrouenstellings vergelyk word, indien laasgenoemde wel in die toekoms beter resultate vertoon.

Die gevolgtrekking is dat dit riskant kan wees om op 'n enkele leksikografiese bron staat te maak vir só 'n komplekse proses. 'n Tweetalige woordeboek is in die eerste plek geskep vir 'n menslike gebruiker en nie as databron vir die skep van nuwe verwysingsbronne waar die struktuur heelwat verskil nie. Die *Groot Woordeboek* voldoen dus nie aan die vereistes wat deur die metodes gestel word nie. Die metodologie behoort moontlik beter te werk indien ander bronne soos 'n tesourus en 'n eentalige woordeboek, wat verdere sinonieme en betekenisonderskeidings gee, betrek en gekombineer kan word.

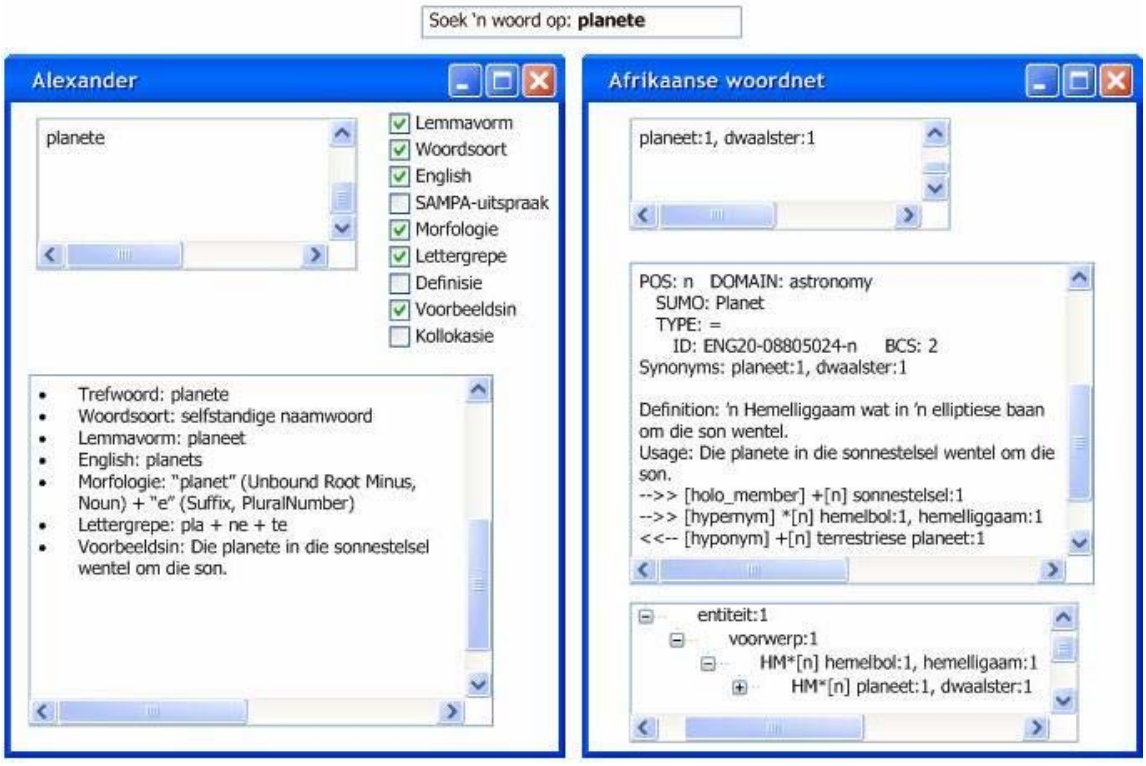
6. Integrasie met ander bronne

Die Afrikaanse woordnet bestaan tans uit 5 041 sinoniemstelle en definisies en is met die hand gebou. Elkeen van hierdie sinoniemstelle is aan Engelse ekwivalente gekoppel. Die woordnet is dus, soos voorheen genoem, reeds outomaties geïntegreer met die *Princeton WordNet*, omdat laasgenoemde as die bronwoordnet dien. Dit word moontlik gemaak deur middel van die XML-struktuur wat deur 'n databasis- en woordnetontwikkelingsprogram, *DEBVisDic*, gelees word (vgl. Horák *et al*., 2003). Enige ander woordnet wat by hierdie struktuur aansluit, kan aan enige ander woordnet gekoppel word. Dit is dus tegnies moontlik om enige sinoniemstel in enige woordnet te kies en die ekwivalente in ander tale op te soek. Die invoer van semantiese verhoudings in 'n woordnet is meestal relatief eenvoudig as gevolg van die struktuur. As gevolg van sy grootte dien die *Princeton WordNet* as middelskakel vir die meeste woordnette in meertalige woordnetdatabasisse. Die lys van Engelse konsepte in hierdie funksie staan bekend as die *Inter-lingual-Index* (*ILI*). Moontlike toekomstige woordnette vir ander Suid-Afrikaanse tale sal daarby baat om hierdie benadering te volg, aangesien dit die integrasie sal verhoog.

Daar word ook beplan om die Afrikaanse woordnet te integreer met 'n ander Afrikaanse leksikale databasis wat ook by die Noordwes-Universiteit ontwikkel word, naamlik *ALEXANDER* (*Afrikaans lexicon and annotated database for engineering and research*). Hierdie bron gaan gedetailleerde talige inligting oor Afrikaanse woorde bevat, onder andere morfologiese analises, woordsoortetikette, *SAMPA*-uit-

spraak, lettergreepverdeling, samestellingsinligting, lemmavorms, definisies, voorbeeldsinne, Engelse vertalings en kollokasies. ’n *SQL*-databasis word gebou met ’n koppelvlak wat die gebruiker in staat sal stel om inligting oor ’n woord op te soek in sowel *ALEXANDER* as die Afrikaanse woordnet, wat in verskillende vensters vertoon sal word. Soos in *DEBVisDic*, sal daar ook tussen die woordnet se sinoniemstelle beweeg kan word volgens semantiese verhoudings, terwyl die relevante data se trefwoordekwivalent in *ALEXANDER* se venster weerspieël sal word. Die woordnet se trefwoorde is egter slegs in lemmavorm, dus moet ’n trefwoord in verboë vorm deur ’n lemmatiseerder omgeskakel word sodat die relevante woord vertoon kan word. Die koppelvlak is generies: ’n mens sal kan kies watter inligting vertoon moet word en watter nie, byvoorbeeld deur middel van merkblokkies. Die blaaier sal toeganklik wees op ’n webblad. Ontwikkeling het reeds begin en ’n beta- of finale weergawe van ’n geïntegreerde *ALEXANDER* sal teen die einde van 2008 gereed wees. Figuur 4 is ’n voorstelling van hoe dit moontlik kan lyk.

Figuur 4: Voorstelling van ’n generiese koppelvlak om woorde in sowel *ALEXANDER* as die Afrikaanse woordnet op te soek



Woordnette vir ander Suid-Afrikaanse tale word reeds ontwikkel by die Universiteit van Suid-Afrika onder leiding van professor Sonja Bosch (Le Roux *et al*., 2007). Die struktuur maak dit moontlik vir elkeen van hierdie woordnette om by mekaar aan te sluit.

7. Slot

In sy huidige formaat word die Afrikaanse woordnet nog slegs as 'n kernwoordnet beskou. Indien dit later uitgebrei word, sal groot baat gevind word by die toepassing van outomatiese metodes, aangesien die handmatige bou van 'n woordnet 'n baie stadige proses is. Hierdie artikel dui slegs die eerste stap in daardie rigting aan. Die integrasie van die woordnet met ander bronne is nog 'n belangrike stap in die ontwikkeling van taalhulpbronne vir Afrikaans en uiteindelik vir die bevordering van mensetaaltegnologie in Suid-Afrika.

Geraadpleegde bronne

ATSERIAS, J., CLIMENT, S., FARRERES, X., RIGAU, G. & RODRIGUEZ, H. 1997. Combining multiple methods for the automatic construction of multilingual wordnets. (*In* Proceedings of the International Conference: Recent Advances on Natural Language Processing (RANLP). Bulgaria: Tzigov Chark. p. 143-149.)

BARBU, E. & BARBU, M.V. 2005. Automatic building of wordnets. (*In* Proceedings of the International Conference: Recent Advances on Natural Language Processing (RANLP). Bulgaria: Borovets. https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2005/RANLP/papers/89_barbu.pdf Date of access: 2 Apr. 2008.

BUSCALDI, D., ROSSO, P. & ARNAL, E.M. 2004. WordNet as a geographical information resource. (*In* Sojka, P., Choi, K., Fellbaum, C., Vossen, P., *eds*. GWC Proceedings. Brno, Czech Republic: Masaryk University. p. 37-42.)

DU PLESSIS, M., *red*. 2005. Pharos Afrikaans-Engels/English-Afrikaans-woordeboek/dictionary. Kaapstad: Pharos.

EKSTEEN, L.C. 1997. Groot Woordeboek: Afrikaans-Engels, Engels-Afrikaans. Kaapstad: Pharos.

EVENS, M., *ed*. 1988. Relational models of the lexicon. Cambridge: Cambridge University Press.

FELLBAUM, C., *ed*. 1998. WordNet: an electronic lexical database. Cambridge: MIT.

FELLBAUM, C., PALMER, M., HOA, T.D, DELFS, L. & WOLF, S. 2001. Manual and automatic semantic annotation with WordNet. (*In* Proceedings of the NAACL: workshop on WordNet and other lexical resources. Pittsburgh: Carnegie Mellon University.)

GLOBAL WORDNET ASSOCIATION. s.a. Base concepts. http://www.globalwordnet.org/gwa/gwa_base_concepts.htm Date of access: 2 Apr. 2008.

HORÁK, A., PALA, K., RAMBOUSEK, A. & POVOLNÝ, M. 2003. DEBVisDic: first version of new client-server WordNet browsing and editing tool. (*In* Proceedings of the 2nd International WordNet Conference (GWC 2004), Brno, Czech Republic. p. 136-141.)

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. s.a. http://www.ieee.org/ Date of access: 2 Apr. 2008.

KIM, Y., CHANG, J. & ZHANG, B. 2002. Target word selection using WordNet and data-driven models in machine translation. (*In* PRICAI 2002: Trends in Artificial Intelligence: 7th Pacific Rim International Conference on Artificial Intelligence, Tokyo, Japan, August 18-22, 2002. Proceedings. Heidelberg: Springer. p. 467-471.)

LE ROUX, J., MOROPA, K., BOSCH, S. & FELLBAUM, C. 2007. Introducing the African languages wordnet. (*In* Tanács, A., Csendes, D., Vincze, V., Fellbaum, C. & Vossen, P., *eds.* Proceedings of the Fourth Global WordNet Conference, Szeged, Hungary, January 22-25, 2008. Szeged: University of Szeged, Department of Informatics. p. 269-280.)

LEE, D., OH, J., CHOE, H. & CHOI, J. 2005. Research on processing a multiple nominative case construction in Korean-English: machine translation by using WordNet. (*In* Sojka, P., Choi, K., Fellbaum, C., Vossen, P., *eds*. GWC Proceedings. Brno, Czech Republic: Masaryk University. p. 207-210.)

MANDALA, R., TOKUNAGA, T. & TANAKA, H. 1999. Complementing WordNet with Roget's and corpus-based thesauri for information retrieval. Proceedings of EACL. http://delivery.acm.org/10.1145/980000/977049/p94-mandala.pdf Date of access: 2 Apr. 2008.

MEGAPUTER INTELLIGENCE. s.a. http://www.megaputer.com/ Date of access: 2 Apr. 2008.

MEL'ČUK, I.A. 1996. Lexical functions: a tool for the description of lexical relations in a lexicon. (*In* Wanner, Leo. Lexical functions in lexicography and natural language processing. Amsterdam: Benjamins. p. 37-102.)

MORATO, J., MARZAL, M., LLORÉNS, J. & MOREIRO, J. 2004. WordNet applications. (*In* Sojka, P., Pala, K., Smrž, P., Fellbaum, C. & Vossen, P., *eds*. GWC Proceedings. Brno, Czech Republic: Masaryk University. p. 273-274.) www.fi.muni.cz/gwc2004/proc/105.pdf Date of access: 2 Apr. 2008.

PEASE, A., NILES, I., & LI, J. 2002. The suggested upper merged ontology: a large ontology for the semantic web and its applications. Working notes of the AAAI-2002 workshop on ontologies and the semantic web. Edmonton, Canada.

PHAROS WOORDEBOEKE DICTIONARIES. s.a. http://www.pharosaanlyn.co.za Datum van gebruik: 2 Apr. 2008.

PIANTA, E., BENTIVOGLI, L. & GIRARDI, C. 2002. MultiWordNet: developing an aligned multilingual database. (*In* Proceedings of the 1st International WordNet Conference, Mysore, India, p. 293-302.)

TUFIŞ, D., CRISTEA, D. & STAMOU, S. 2004. BalkaNet: aims, methods, results and perspectives: a general overview. *Romanian journal of information science and technology*, 7(1-2):9-43.

VAN SCHALKWYK, D.J., *red*. 2005. Verklarende Woordeboek van die Afrikaanse Taal (CD-ROM). Buro van die WAT.

VOSSEN, P. 1999. EuroWordNet general document. Amsterdam: University of Amsterdam.

VOSSEN, P., BLOKSMA, L., BOERSMA, P. 1999. The Dutch WordNet. http://www.vossen.info/docs/1999/DutchWordNet.pdf Date of access: 2 Apr. 2008.

WORDNET: A LEXICAL DATABASE FOR THE ENGLISH LANGUAGE – WORDNET STATISTICS. s.a. http://wordnet.princeton.edu/man/wnstats.7WN Date of access: 2 Apr. 2008.

Kernbegrippe:

Afrikaanse woordnet
ALEXANDER-databasis
woordnetintegrasie
woordnetkonstruksie

Key concepts:

Afrikaans WordNet
ALEXANDER database
wordnet construction
wordnet integration