http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Optimalisering van gratis elektroniese/aanlyn hulpbronne vir woordeboeksamestelling — 'n drietalige woordeboekeksperiment

Sonja E. Bosch, Departement Afrikatale,
Universiteit van Suid-Afrika, Pretoria, Suid-Afrika
(*seb@hbosch.com*)
ORCID: *https://orcid.org/0000-0002-9800-5971*

Marissa Griesel, Departement Afrikatale,
Universiteit van Suid-Afrika, Pretoria, Suid-Afrika
(*griesm@unisa.ac.za*)
ORCID: *https://orcid.org/0000-0003-1309-0212*
en
Elsabé Taljard, Departement Afrikatale,
Universiteit van Pretoria, Pretoria, Suid-Afrika
(*elsabe.taljard@up.ac.za*)
ORCID: *https://orcid.org/0000-0002-4507-1633*

Opsomming: Die beskikbaarheid van meertalige woordeboeke is deurslaggewend, nie slegs vir direkte teikengebruikers nie, maar ook vir indirekte teikengebruikers soos menslike taaltegnoloë, veral in die geval van tale met skaars hulpbronne, soos byvoorbeeld Venda. In hierdie artikel word die optimale benutting van gratis elektroniese/aanlyn hulpbronne vir die samestelling van 'n drietalige e-woordeboek vir Venda, Engels en Afrikaans ondersoek. Ons benadering is gebaseer op 'n eksperiment waarin die samestellingsproses so ver moontlik geoutomatiseer is om besparing in terme van tyd en mensekrag teweeg te bring. Engels word as 'n brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Die algemene bevindinge is dat daar sekere beperkings te wagte kan wees in so 'n semi-outomatiese proses wat wel 'n sekere mate van menslike intervensie verg. Hoewel die saamgestelde e-woordeboek nie as 'n finale produk beskou kan word nie, bied die woordeboeksamestellingsprogram *Lexonomy*, wat vanweë sy aanpasbaarheid en maklike uitleg suksesvol in hierdie studie gebruik is, die geleentheid vir menslike insette om die nodige aanpassings op 'n gebruikersvriendelike wyse te doen. Die geformuleerde konsepvoorstel is nuttig vir die skep van meertalige aanlyn woordeboeke, saamgestel met behulp van beskikbare aanlyn of elektroniese hulpbronne. Die resulterende drietalige woordeboek is aanlyn beskikbaar as bewys van die konsep waarop verdere werk kan bou. Die feit dat die databasis onderliggend aan die woordeboek beskikbaar is in 'n masjienleesbare formaat, naamlik XML, is belangrik vir indirekte

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

teikengebruikers vir hergebruik om elektroniese hulpbronne te ontwikkel, veral vir hulpbronarm tale.

Sleutelwoorde: WOORDEBOEKSAMESTELLING, VENDA–ENGELS–AFRIKAANS, DRIETALIGE WOORDEBOEK, ELEKTRONIESE/AANLYN HULPBRONNE, MASJIENVERTAALSISTEME, *LEXONOMY*, KORPUSSOEKTOG, TEIKENGEBRUIKERS

Abstract: Optimization of Free Online/Electronic Resources for Dictionary Compilation — A Trilingual Dictionary Experiment. The availability of multilingual dictionaries is crucial, not only for direct target users, but also for indirect target users, especially in the case of languages with scarce resources such as Venda. This article explores the optimal use of free electronic/online resources for compiling a trilingual e-dictionary for Venda, English and Afrikaans. Our approach is based on an experiment in which the compilation process was automated as far as possible to achieve savings in terms of time and manpower. English is used as a bridge for the translation between the source language, Venda, and the target language, Afrikaans. The general finding is that certain limitations can be expected in such a semi-automated process that requires a certain amount of human intervention. Although the composite e-dictionary cannot be considered a final product, the dictionary compilation program *Lexonomy*, which has been used successfully in this study due to its adaptability and easy layout, provides the opportunity for human input to make the necessary adaptations in a user-friendly manner. The proposed concept is useful for creating multilingual online dictionaries, compiled using available online or electronic resources. The resulting trilingual dictionary is available online as proof of concept on which further work can build. The fact that the database underlying the dictionary is available in a machine-readable format, namely XML, is important for indirect target users for reuse to develop electronic resources, especially for resource-scarce languages.

Keywords: DICTIONARY COMPILATION, VENDA–ENGLISH–AFRIKAANS, TRILINGUAL DICTIONARY, ELECTRONIC/ONLINE RESOURCES, MACHINE TRANSLATION SYSTEMS, *LEXONOMY*, CORPUS SEARCH, TARGET USERS

1. Inleiding

Woordeboeksamestelling vir tale met skaars hulpbronne is 'n arbeidsintensiewe taak. In hierdie artikel word die optimale benutting van gratis elektroniese/aanlyn hulpbronne vir die saamstel van 'n drietalige woordeboek ondersoek en bespreek. Die betrokke tale is Venda, Engels en Afrikaans. Die doelstellings van hierdie studie is eerstens om met 'n unieke benadering te eksperimenteer waar Engels as 'n brug vir die (semi-outomatiese) vertaling tussen die brontaal, Venda en die doeltaal, Afrikaans, gebruik word. 'n Verdere doelstelling is die formulering van 'n konsepvoorstel wat gebruik kan word vir die daarstel van 'n drietalige aanlyn woordeboek, saamgestel met behulp van gratis hulpbronne. Ons ondersoek in die derde plek die mate waartoe die samestellingsproses geoutomatiseer kan word, aangesien 'n (semi-) outomatiese benadering 'n besparing in terme van tyd en mensekrag teweeg kan bring. Die semi-outo-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

matiese benadering wat in hierdie studie voorgestel en beskryf word, verhoog ook die toeganklikheid tussen minderheidstale (Venda en Afrikaans). Die resulterende drietalige woordeboek sal verder met behulp van gratis hulpbronne ook aanlyn beskikbaar gestel word as 'n bewys van die konsep waarop verdere werk kan bou.

In die volgende afdeling word 'n oorsig gegee van bestaande woordeboeke met Venda en Afrikaans as taalpaar, waarna twee tipes potensiële teikengebruikers in Afdeling 3 geïdentifiseer word. Afdeling 4 beskryf die beskikbare gratis aanlyn/elektroniese hulpbronne waarop die eksperiment (soos beskryf in Afdeling 5) gebaseer is. Dit word gevolg deur 'n evaluering van die eksperiment in Afdeling 6 en laastens, in Afdeling 7, voorstelle vir toekomstige werk.

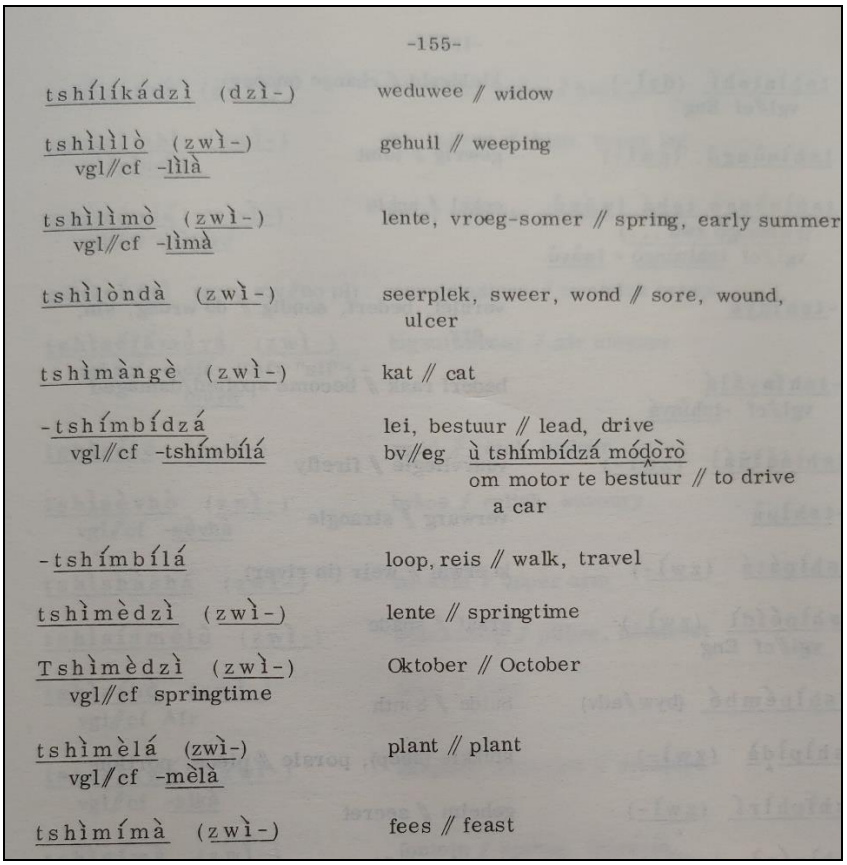
2. Bestaande woordeboeke met Venda en Afrikaans as taalpaar

Die enigste woordeboek waarin Afrikaans en Venda as taalpaar voorkom, is die *Drietalige Elementêre Woordeboek* (DEW) deur Wentzel en Muloiwa (1976). In die voorwerk (ibid: i) van die woordeboek word die doel van die woordeboek soos volg verwoord:

> Hierdie elementêre drietalige woordeboek is bedoel om — al is dit totdat 'n meer volledige uitgawe kan volg — in 'n dringende behoefte te voorsien. Daar is naamlik geen Venda-woordeboek beskikbaar nie en dit is bykans onmoontlik om byvoorbeeld die Vendataal te probeer aanleer sonder 'n beskikbare woordeboek .... Die enigste woordeboek wat nog ooit saamgestel is, is die *Tshivenḓa–English Dictionary* deur NJ van Warmelo (1937). Hierdie boek is egter reeds vir baie jare uit druk en hoewel dit 'n baie betroubare naslaanwerk is, is dit nie vryelik beskikbaar vir studente nie.

Die DEW bestaan uit drie dele. In die eerste deel is die taal van lemmatisering Venda, met Afrikaanse en Engelse ekwivalente, deel twee bestaan uit 'n Afrikaans–Venda-afdeling, gevolg deur deel drie, die Engels–Venda-afdeling. Die woordeboek het 'n raamstruktuur bestaande uit 'n voorwerk, gevolg deur die drie alfabetiese lemmalyste. Die lemmatiseringstrategie in deel een is woordgebonde, soos meestal die geval is vir disjunktiefgeskrewe Afrikatale. Die ordening is nie streng alfabeties nie — waar lemmas met konsonante begin wat 'n diakritiese teken gebruik om die Romeinse alfabet aan te vul, word hierdie lemmas in 'n aparte alfabetiese strek geplaas. Die dentale simbole ḓ, ḽ, ṋ en ṱ gaan die gewone d, l, n en t vooraf, terwyl die velêre ṅ ná gewone n volg. Die artikelstruktuur is relatief eenvoudig. Woordkategorieë word slegs op implisiete wyse aangedui — in die geval van naamwoorde word die meervoud aangedui deur die meervoudsprefiks in hakies ná die lemma te verstrek, gevolg deur die Afrikaanse en Engelse vertaalekwivalente. Vergelyk byvoorbeeld die lemma *tshimedzi* 'lente' in Figuur 1:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)



-155-

tshílíkádzì (dzì-) — weduwee // widow

tshìlìlò (zwì-) — gehuil // weeping
vgl//cf -lìlà

tshìlìmò (zwì-) — lente, vroeg-somer // spring, early summer
vgl//cf -lìmà

tshìlòndà (zwì-) — seerplek, sweer, wond // sore, wound, ulcer

tshìmàngè (zwì-) — kat // cat

-tshímbídzá — lei, bestuur // lead, drive
vgl//cf -tshímbílá — bv//eg ù tshímbídzá módòrò — om motor te bestuur // to drive a car

-tshímbílá — loop, reis // walk, travel

tshìmèdzì (zwì-) — lente // springtime

Tshìmèdzì (zwì-) — Oktober // October
vgl//cf springtime

tshìmèlá (zwì-) — plant // plant
vgl//cf -mèlà

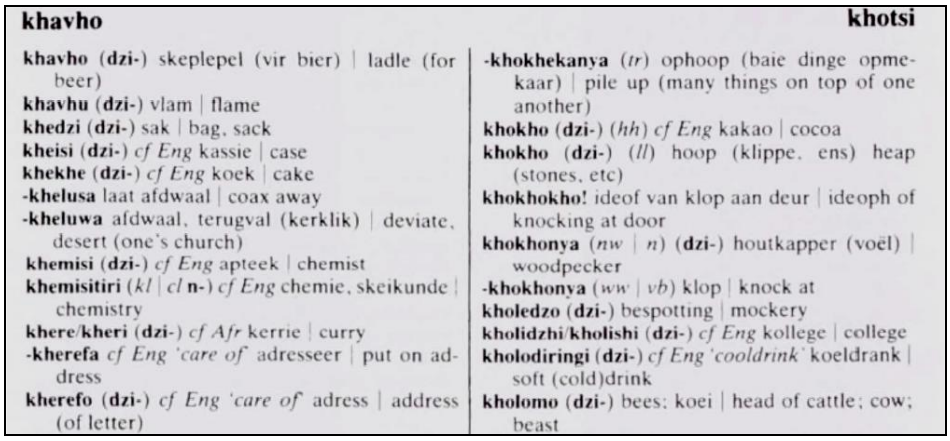
tshìmímà (zwì-) — fees // feast

Figuur 1: Uittreksel uit die *Drietalige Elementêre Woordeboek* deur Wentzel en Muloiwa (1976: 155)

In gevalle waar meervoudsvorming onreëlmatig is of waar fonologiese verandering plaasvind, word die volledige meervoudsvorm verstrek — vergelyk byvoorbeeld *danda* 'paal', meervoud *matanda*. In die geval van deverbatiewe naamwoorde word 'n kruisverwysing na die werkwoordstam wat as basis vir die deverbatief dien, aangegee, sien byvoorbeeld *tshimela* 'plant (naamwoord)' met kruisverwysing na *-mela* 'plant (werkwoordstam)' in Figuur 1 hierbo. By werkwoorde as lemmas word die werkwoordstatus aangedui deur 'n koppelteken wat die stam voorafgaan, vergelyk *-tshimbila* 'loop'. Afgeleide werkwoordstamme word gelemmatiseer en behandel, maar 'n kruisverwysing na die basisvorm word ook verstrek, byvoorbeeld *-tshimbidza* 'lei, bestuur' met kruisverwysing na *-tshimbila* 'loop'. Toon word in alle gevalle aangedui. Etimologiese inligting word in die geval van leenwoorde uit Afrikaans of Engels verskaf, en gebruiksvoorbeelde word op skynbaar lukrake wyse verstrek.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

In 1982 verskyn 'n verbeterde uitgawe van hierdie woordeboek wat in 2009 'n sesde druk beleef. Waar die 1976-weergawe op studente gemik was wat die toenmalige Spesiale Kursus in Venda aan UNISA gevolg het, word daar in die voorwoord tot die 1982-weergawe aangedui dat 'n breër spektrum gebruikers die teiken van dié uitgawe is. Dit word beplan as die eerste in 'n reeks van drie woordeboeke: 'n praktiese klein woordeboek, wat beplanningsgewys gevolg sou word deur 'n omvattende twee- of drietalige woordeboek wat onder andere ook idiomatiese taalgebruik en die aanduiding van toon sal insluit, en 'n naslaanwerk soortgelyk aan dié van N.J. van Warmelo se *Tshivenḓa–English Dictionary* (1937). In die 1982-weergawe word die aantal inskrywings verdubbel, maar voorbeelde word beperk tot enkele gevalle waar dit as onontbeerlik beskou word. Daar word ook weggedoen met die aanduiding van toon, behalwe in gevalle waar toon 'n woordonderskeidende funksie het, vergelyk byvoorbeeld *khokho* 'kakao' en *khokho* 'hoop klippe' in Figuur 2 hieronder. Toon word ook nie ortografies aangedui nie, maar met behulp van die letters *h* = hoog en *l* = laag. Kruisverwysings na basisvorms in geval van afgeleide werkwoordstamme en deverbatiewe word ook weggelaat. Enkele gebruiksnotas word in hakies verstrek, bv. "*evho*! O nee! (slegs vroue)" (Wentzel en Muloiwa 1982: 14). Vergelyk Figuur 2 vir 'n uittreksel uit die verbeterde weergawe:



khavho **khotsi**

khavho (**dzi-**) skeplepel (vir bier) | ladle (for beer)
khavhu (**dzi-**) vlam | flame
khedzi (**dzi-**) sak | bag, sack
kheisi (**dzi-**) *cf Eng* kassie | case
khekhe (**dzi-**) *cf Eng* koek | cake
-khelusa laat afdwaal | coax away
-kheluwa afdwaal, terugval (kerklik) | deviate, desert (one's church)
khemisi (**dzi-**) *cf Eng* apteek | chemist
khemisitiri (*kl* | *cl* **n-**) *cf Eng* chemie, skeikunde | chemistry
khere/kheri (**dzi-**) *cf Afr* kerrie | curry
-kherefa *cf Eng 'care of'* adresseer | put on address
kherefo (**dzi-**) *cf Eng 'care of'* adress | address (of letter)
-khokhekanya (*tr*) ophoop (baie dinge opmekaar) | pile up (many things on top of one another)
khokho (**dzi-**) (*hh*) *cf Eng* kakao | cocoa
khokho (**dzi-**) (*ll*) hoop (klippe, ens) heap (stones, etc)
khokhokho! ideof van klop aan deur | ideoph of knocking at door
khokhonya (*nw* | *n*) (**dzi-**) houtkapper (voël) | woodpecker
-khokhonya (*ww* | *vb*) klop | knock at
kholedzo (**dzi-**) bespotting | mockery
kholidzhi/kholishi (**dzi-**) *cf Eng* kollege | college
kholodiringi (**dzi-**) *cf Eng 'cooldrink'* koeldrank | soft (cold)drink
kholomo (**dzi-**) bees; koei | head of cattle; cow; beast

Figuur 2: Uittreksel uit die *Verbeterde Drietalige Woordeboek* deur Wentzel en Muloiwa (1982: 24)

Sedert die laaste druk in 2009 is hierdie woordeboek egter uit druk. Dit is waarskynlik nie kommersieel haalbaar om dit te herdruk nie, maar dit beteken nie dat die behoefte aan so 'n woordeboek nie meer ter sake is nie. Dit is daarom die moeite werd om ondersoek in te stel na die moontlikheid om met weinig kostes 'n Venda–Afrikaans woordeboek saam te stel en gratis aanlyn aan gebruikers beskikbaar te stel.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

3. Potensiële teikengebruikers

Vir die doel van hierdie bespreking onderskei ons tussen twee tipes gebruikers, te wete direkte en indirekte teikengebruikers. Direkte teikengebruikers is diegene wat die woordeboek gebruik om vertaalekwivalente van byvoorbeeld Venda-lemmas in Afrikaans of Engels te vind, of enige variasie van hierdie konfigurasie. So 'n basiese vertaalwoordeboek sal van nut wees vir Afrikaans- of Engelssprekendes wat Venda wil aanleer, of selfs Vendasprekers wat Afrikaans of Engels aanleer. Onderwysstudente sou besonder baat vind by so 'n woordeboek. In die Departement van Hoër Onderwys en Opleiding se hersiene beleid (2015) vir die minimum vereistes vir onderwyskwalifikasies word gestipuleer dat alle afgestudeerde onderwysstudente vaardig moet wees in ten minste een van Suid-Afrika se amptelike tale as 'n taal van onderrig en leer (TvOL). Verder moet elke student kommunikatief vaardig wees in ten minste een ander amptelike Afrikataal. In gevalle waar die taal van onderrig en leer Afrikaans of Engels is, moet die taal van kommunikatiewe vaardigheid 'n Afrikataal wees. Dit is dus moontlik dat 'n student wat Afrikaans as TvOL aanbied, Venda as taal van kommunikatiewe vaardigheid kan kies. 'n Basiese vertaalwoordeboek sal in so 'n geval 'n bruikbare hulpbron wees. Aangesien die woordeboek in hierdie eksperiment nie 'n baie uitgebreide bewerking van lemmas in die vooruitsig stel nie, teiken dit die tipiese 'on the fly'-gebruiker — iemand wat binne 'n bepaalde gebruiksituasie bloot na 'n vertaalekwivalent op soek is. Die beskikbaarheid van so 'n aanlyn hulpbron kan ook die elektroniese voetspoor van 'n minderheidstaal soos Venda vergroot en sodoende bydra tot die status van dié taal as taal van hoër funksies.

Indirekte teikengebruikers stel belang in die databasis wat die woordeboek onderlê, veral as die data in 'n masjienleesbare formaat soos die internasionaal erkende *Extensible Markup Language*, oftewel XML[1], beskikbaar is. In die ontwikkeling van elektroniese hulpbronne, veral vir hulpbronarm tale, is dit belangrik dat hierdie bronne beskikbaar gestel word vir hergebruik. Sulke hulpbronne vorm die basis vir talle Menslike Taaltegnologietoepassings, soos byvoorbeeld masjienvertaling. Die Autshumato-projek[2] is die enigste grootskaalse poging om masjienvertalingsisteme en -hulpbronne vir Suid-Afrikaanse tale daar te stel (cf. McKellar en Groenewald 2012) en bied tans slegs beperkte hulpbronne in die vorm van parallelle korpora en masjienvertaalgeheues vir sekere Suid-Afrikaanse taalpare gratis aan. Afrikaans en Venda as taalpaar word tans nie in hierdie projek gedek nie. Ṋemuṱamvuni (2018) en Moors et al. (2018) bevestig dat die nodige hulpbronne om selfs masjiengesteunde menslike vertaling (oftewel '*machine aided human translation*' soos deur Sager (1994: 326) gedefinieer) met Venda as doeltaal te ontwikkel, nog nie resultate lewer wat met internasionale standaarde vergelyk kan word nie en dat hierdie agterstand grootliks toegeskryf kan word aan 'n tekort aan masjienleesbare hulpbronne.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

4. Beskikbare gratis aanlyn/elektroniese hulpbronne

4.1 Venda: CBOLD (Murphy 1997) — Venda–Engels Woordeboek

In 1994 loods Larry Hyman en John Lowe 'n projek wat ten doel het om 'n aanlyn leksikografiese databasis vir navorsers en leksikograwe daar te stel. As deel van die *Comparative Bantu OnLine Dictionary*-projek (CBOLD) word 'n verskeidenheid hulpbronne in 'n heterogene digitale formaat beskikbaar gestel. 'n Aantal Bantoetaalwoordeboeke word onder 'n oop lisensie aangebied, soos vermeld in die 'Bantuists' Manifesto' wat op die webblad verskyn. Tussen 1994 en 2000 is 'n groot aantal Bantoe-woordeboeke deur CBOLD gedigitiseer en via die projekwebblad vir gebruik en toepassings beskikbaar gestel. Die CBOLD-woordeboeke word in inkonsekwente datastrukture en skemas en in 'n verskeidenheid van formate aangebied (sien ook Eckart et al. 2019: 17.4-17.5). Uiteraard kan hierdie skematiese en tegniese heterogeniteit nie sonder meer vir verdere toepassings gebruik word nie. Transformasie- en kwaliteitsversekeringsmaatreëls is nodig alvorens hierdie waardevolle leksikale databron aktief gebruik kan word.

Een van die beskikbare woordeboeke is 'n Venda-woordeboek (Murphy 1997) wat in die formaat van 'n gewone tekslêer is. Die volgende datavelde word in die woordeboek onderskei: vir naamwoorde word die toonpatroon, die woordklas, die klasnommer waartoe die naamwoord behoort en die Engelse vertaalekwivalent aangedui; vir werkwoorde word die toon van die werkwoordstam aangedui, gevolg deur die woordklas en die vertaalekwivalent in Engels. Hierdie woordeboek vorm die basis van die eksperiment wat in Afdeling 5 beskryf word.

4.2 Gratis Engels–Afrikaanse masjienvertaalsisteme

Masjienvertaling is reeds in die 1940's met behulp van ponskaarte gedoen. Aanvanklik was die kwaliteit van masjienvertaling swak, en selfs tans is sodanige vertaling steeds nie perfek nie, maar soos Groves en Mundt (2015: 113) opmerk, kan die ontwikkeling van kunsmatige intelligensie tot die ontwikkeling van gesofistikeerde vertaalsisteme aanleiding gee. Dit maak daarom sin om die potensiaal van sodanige sisteme te ondersoek.

In hierdie eksperiment word Engels as brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Gratis beskikbare masjienvertaalsisteme wat vir die vertaling van die Engelse vertaalekwivalente na Afrikaans gebruik sou kon word, is die volgende:

— Google Translate (https://translate.google.com/)

Google Translate is 'n veeltalige neurale masjienvertaaldiens wat deur Google ontwikkel is om teks, dokumente en webwerwe te vertaal. Dit is 'n statisties-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

gebaseerde sisteem, wat impliseer dat die waarskynlikheid van verskeie korrekte vertaalmoontlikhede bereken word, eerder as wat die sisteem op 'n woord-vir-woord vertaling afgestem is (Groves en Mundt 2015: 113). Tans ondersteun Google Translate 109 tale op verskillende vlakke. Dit bied 'n webwerf-koppelvlak, 'n mobiele toepassing vir Android en iOS en 'n toepassingsprogrammeerkoppelvlak wat ontwikkelaars help om uitbreidings vir webblaaiers ("browsers") en programmatuurtoepassings te bou[3].

— Bing Microsoft® Translator (https://www.bing.com/translator/)

Microsoft® Translator is 'n meertalige wolkvertalingsdiens vir masjienvertalings wat deur Microsoft® gelewer word. Die diens ondersteun tans 87 taalstelsels. Microsoft® Translator bied ook dienste vir teksvertaling via die Translator Text API, wat wissel van 'n gratis vlak wat twee miljoen karakters per maand ondersteun tot betaalde vlakke wat miljarde karakters per maand ondersteun[4].

— Yandex.Translate (https://translate.yandex.com/)

Yandex.Translate, 'n webdiens wat deur die soekenjin en webportaal Yandex verskaf word, is bedoel vir die vertaling van teks of webblaaie. Vertalings is tans beskikbaar in 98 tale. Die stelsel volg 'n hibriede benadering wat statistiese masjienvertaling en neurale masjienvertalingsmodelle kombineer. 'n Woordeboek van enkelwoordvertalings word op grond van die ontleding van miljoene vertaalde tekste gebou. Om die teks te vertaal, vergelyk die algoritme dit eers met 'n databasis van woorde en vergelyk dan die teks met die basistaalmodelle en probeer die betekenis van 'n uitdrukking in die konteks van die teks bepaal[5].

— english-afrikaans.co.za (https://english-afrikaans.co.za/)

Daar is geen verdere beskrywing van hierdie vertaalsisteem beskikbaar nie, en ongelukkig is die vertaalsisteem as sodanig ook aanlyn verwyder sedert die vertalings vir hierdie projek in Julie 2021 afgehandel is.

4.3 Gratis Afrikaanse speltoetser: WSpel

WSpel[6] is tans die omvattendste Afrikaanse speltoetser wat gratis beskikbaar is. Dit voldoen bowendien aan die spelwyse van die jongste *Afrikaanse Woordelys en Spelreëls* (AWS) 2009. Die woordelys bestaan uit ietwat meer as 526,000 woorde en daar is duisende inskrywings wat in Microsoft® Word se AutoCorrect-funksie gebruik kan word. Vir hierdie eksperiment het ons WSpel 15 afgelaai en geïnstalleer.

4.4 Gratis korpussoektogprogramme

Korpussoektogprogramme is onmisbaar vir moderne woordeboeksamestelling,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

aangesien dit gebruik word om frekwensie-inligting uit 'n korpus te onttrek wat belangrik is vir die saamstel van 'n lemmalys of die aanvulling van 'n bestaande een. In die latere fases van samestelling kan dit onder andere gebruik word vir betekenisonderskeidings tussen polisemiese lemmas en vir die identifisering van gebruiksvoorbeelde en frekwente kollokasies. Vir tale soos Afrikaans, en veral Venda, wat van diakritiese tekens in die ortografie gebruik maak, is dit belangrik dat korpussoektogprogramme hierdie tekens korrek kan hanteer en weergee in die resultate van 'n korpussoektog. Korpussoektogprogramme moet verder die oplaai van eie korpora en die aflaai van soekresultate ondersteun. Vir die doel van hierdie artikel het ons 'n beperkte eksperiment met beide Venda- en Afrikaanse tekste uitgevoer en gevind dat al drie programme hieronder gelys aan bogenoemde vereistes voldoen en dus suksesvol vir korpussoektogte gebruik kan word.

— AntConc: https://www.laurenceanthony.net/software/antconc/
— LancsBox: http://corpora.lancs.ac.uk/lancsbox/
— Voyant Tools: https://voyant-tools.org/

4.5 Venda-korpus

Die gebruik van korpora is reeds standaardpraktyk in moderne woordeboeksamestelling. Korpusdata word gebruik om lemmalyste op te stel, frekwensie-inligting te bekom en is 'n potensiële bron van gebruiksvoorbeelde. 'n Tipiese korpussoektog na sleutelwoorde in konteks ('n sogenaamde KWIC oftewel "Keyword in Context"-soektog, Afrikaans SWIK-soektog) met konkordansielyne as resultaat, is 'n onmisbare bron vir die identifisering van alle moontlike betekenisse van 'n bepaalde lemma. Vir die doel van hierdie eksperiment is 'n Venda-korpus van 1.4 miljoen ortografiese woorde ("tokens") gebruik. Hierdie korpus vorm deel van 'n projek van die *South African Centre for Digital Language Resources* (SADiLaR), wat die daarstel van elektroniese hulpbronne, spesifiek vir die Afrikatale, ten doel het. Dit is 'n rou korpus, sonder enige annotasie van byvoorbeeld woordkategorieë, en bestaan uit gedigitiseerde tekste oor 'n wye verskeidenheid genres, maar sluit nie 'n gesproke komponent in nie. Dit is heel waarskynlik nie gebalanseerd nie, maar om Atkins et al. (1992: 14) aan te haal:

> In our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus — however 'unbalanced' — to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simply because the corpus used cannot be proved to be 'balanced'.

Hierdie opmerking is veral geldig ten opsigte van die Afrikatale, spesifiek vir data-arm tale soos Venda.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

4.6 Gratis program vir woordeboeksamestelling: *Lexonomy*

Die navorsingsprojek genaamd *European Lexicographic Infrastructure* (ELEXIS) het ten doel om die tale van Europa te verbind deur die woordeboeke wat vir hierdie tale bestaan te standaardiseer, digitiseer en met mekaar te vergelyk (ELEXIS 2020: 1):

> More people than ever before use dictonaries. Not so much printed ones, but dictonary data are now built into mobile phones, software, systems and are in use all the time. New and updated dictonaries are badly needed fast (ELEXIS 2020: 1).

Om daardie doel te bereik, het die projek 'n verskeidenheid hulpbronne en gereedskapstelle ontwikkel en gratis vir leksikograwe, ontwikkelaars van Mensliketaaltegnologie (MTT) en taalgebruikers beskikbaar gestel. Een van die mees prominente hulpmiddels is *Lexonomy* — 'n wolkgebaseerde, oopbron woordeboekskrywer en -publiseerder[7].

Die ontwikkelaars van *Lexonomy* wou 'n hulpbron skep wat geen installasie, programmeringskennis of duur kostes dra om 'n basiese aanlyn woordeboek te kan skep nie en só meer gebruikers in staat stel om hul eie woordeboeke op te stel (Měchura 2017: 662). Die platform kan gebruik word om 'n een- of meertalige woordeboek op te stel, te formateer, aan te pas en uiteindelik aanlyn te publiseer. Dit is verder ook moontlik om 'n kollektiewe projek op te stel sodat 'n groep mense saam aan 'n woordeboek kan werk. Die woordeboek bly 'n privaat projek totdat die eienaar dit publiek maak. *Lexonomy* se webblad word dan 'n platform waarop die woordeboek gestoor en gebruik word en van waar dit gedeel word sodat enige gebruiker toegang kan kry.

Alhoewel *Lexonomy* gebruik kan word om 'n nuwe woordeboek van die begin af te ontwikkel en leksikograwe toerus om elke leksikale item stelselmatig van tradisionele datatipes soos die hoofwoord of lemma, woordsoort, definisie en gebruiksvoorbeeld te voorsien, is dit ook moontlik om bestaande data (semi-)outomaties in die sisteem in te voer (Měchura 2017: 664). Die data kan enigiets van 'n eenvoudige woordelys tot 'n uitgebreide veeltalige woordeboek wees, solank dit aan basiese vereistes vir XML formaat voldoen en elke datatipe met die toepaslike merkers geïdentifiseer kan word. As 'n bestaande datastel opgelaai word, kan dit steeds in naredigering deur 'n leksikograaf aangepas/uitgebrei word of volgens 'n nuwe protokol geformateer word. Op dieselfde manier kan 'n woordeboek wat in *Lexonomy* opgestel is, ook in XML formaat afgelaai word sodat dit op 'n ander platform versprei of verder gebruik kan word.

Hierdie twee funksies (die op- en aflaai na en van die XML-databasis) is veral belangrik omdat dit die volhoubare ontwikkeling en gebruik van die resulterende woordeboek verseker. Die XML-formaat kan geredelik deur ander MTT-toepassings hergebruik word omdat die struktuur en merkers vooraf bepaal en in 'n stylgids beskryf word. Elke leksikale item in die woordeboek

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

word volgens dieselfde voorafbepaalde protokol behandel en spesifieke datavelde vir elke item kan dus onttrek word, afhangend van die toepassing waarvoor die data aangewend sal word, byvoorbeeld as afrigtingsdata vir masjienvertaling en inligtingonttrekking, as die basis van 'n woordnet of 'n domeinspesifieke woordeboek. Die XML-struktuur beteken ook dat 'n ontwikkelaar wat nie noodwendig 'n spreker van die taal is nie die datavelde in elke leksikale item kan identifiseer en verder manipuleer vir ander toepassings. So bevorder aanlyn leksikografie ook die ontwikkeling van tale met minder hulpbronne in die MTT-arena.

Lexonomy is al in verskeie soortgelyke projekte aangewend. Stemle et al. (2019: 537-546) beskryf die skep van 'n neologismewoordeboek vir 'n standaardvariant van Duits en hoe *Lexonomy* toegepas word om die woordeboek maklik leesbaar en, belangriker nog, maklik aanpasbaar te maak sodat lemmas gereeld bygevoeg kan word. In hierdie projek word bestaande korpora en woordeboeke ook ingespan om die nuwe woordeboek volgens die sogenaamde "eenkliek-woordeboek paradigma" (oftewel die "one-click dictionary paradigm") saam te stel. Daarin lê natuurlik die grootste verskil tussen die eksperiment wat hier beskryf word en hierdie internasionale projek — vir Venda bestaan daar nog weinig hulpbronne om in die ontwikkelingspyplyn te gebruik.

Bartolomé-Díaz en Frontini (2020: 62-68) gebruik dieselfde komponente om 'n Frans–Spaans tweetalige woordeboek saam te stel, maar fokus hul studie op die kennis en vaardighede wat die samesteller nodig het om so 'n taak suksesvol uit te voer. Die gevolgtrekking is dat *Lexonomy* verder verbeter kan word om byvoorbeeld meer metadata saam met die woordeboek te versamel en om hiperskakels by elke inskrywing toe te laat. Hulle sluit ook by die werk van Jakubíček et al. (2018: 65-68) aan wat op uitdagings in die formatering van woordeboeke volgens streng internasionale standaarde en tydens naredigering wys, maar albei is dit eens dat hierdie uitdagings maklik deur 'n bekwame leksikograaf en versigtige outomatisering van dele van die formateringstappe in *Lexonomy* omseil kan word.

5. Eksperiment

Hierdie afdeling beskryf die stappe wat gevolg is om die effektiwiteit van woordeboeksamestelling met die beskikbare hulpbronne, soos reeds bespreek, te toets. Figuur 6 aan die einde van die afdeling verteenwoordig 'n grafiese voorstelling van die prosedure wat in hierdie eksperiment gevolg is.

5.1 Voorafredigering

Die Engels–Venda woordeboek is eers afgelaai en na 'n eenvoudige sigblad in Microsoft® Excel oorgedra. Deur die data in kolomme te verdeel en 'n unieke ID aan elkeen van die inskrywings toe te ken, kon die navorsingspan maklik

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

besluit watter velde in watter stappe gebruik sou word. Die Engelse vertalings kon byvoorbeeld maklik van die Venda-ekwivalente geskei word om in die masjienvertalingsproses te gebruik.

Vir die doeleindes van hierdie eksperiment is 10% van die woordeboek-inskrywings ewekansig uitgesoek. Geslote woordklasse is egter uitgesluit, dus is 800 ewekansig gekose terme uit die volgende 4 woordsoortklasse ingesluit: naamwoorde, werkwoorde, bywoorde en adjektiewe. Hierdie beginsel is uit masjienleer geleen. Die verdeling van 'n korpus in 3 dele waar 80% vir die afrigting van 'n model gebruik word, 10% vir die evaluering van die model en 10% vir 'n ontwikkelingstel word algemeen aanvaar. Hierdie verdeling word ook in Jurafsky en Martin (2009: 187) aanbeveel wanneer hulle oor evaluering en fout-analise skryf:

> We train our tagger on the training set. Then we use the development test set (also called a dev-test set) to perhaps tune some parameters, and in general decide what the best model is. Once we come up with what we think is the best model, we run it on the (hitherto unseen) test set to see its performance. We might use 80% of our data for training and save 10% each for dev-test and test. Why do we need a development test set distinct from the final test set? Because if we used the final test set to compute performance for all our experiments during our development phase, we would be tuning the various changes and parameters to this set.

5.2 Outomatiese vertaling van Engels na Afrikaans

Tydens die voorafredigering van die Engelse weergawe vir outomatiese vertaling van Engels na Afrikaans is die volgende afkortings outomaties met 'n vind-en-vervang-proses na hul volle vorm verander:

s/t na *something*
s/o na *someone*
w/ na *with*
everything/body na *everything/everybody*

Die Engelse data uit die ontwikkelingstel (oftewel "training set", soos bo beskryf), is hierna outomaties na Afrikaans vertaal met die vier masjienvertaalsisteme wat in Afdeling 4.2 beskryf is. Die resulterende Afrikaanse vertalings is daarna ook in die sigblad gevoeg om nou 'n ontwikkelingstel met een Engelse, een Venda en vier (moontlike) Afrikaanse vertalings in te sluit.

Die volgende stap was om die vertalings te evalueer om sodoende 'n enkele geskikte Afrikaanse vertaling te kies. Hiervoor is menslike intervensie nodig. Vir die handmatige evaluering is punte vir die vertalings deur elke vertaalsisteem toegeken. 'n Perfekte vertaling het twee punte gekry terwyl aan 'n vertaling met bv. 'n spelfout, foutiewe woordorde, foutiewe ortografie, ens. slegs een punt toegeken is. Onbruikbare vertalings het geen punt ontvang nie.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Drie eerstetaalsprekers van Afrikaans het die 800 terme van die ontwikkelingstel geëvalueer en verskille in puntetoekenning met mekaar bespreek om op 'n finale punt te besluit. Afdeling 6 beskryf die evaluering van die masjienvertaalsisteme en die foutanalise breedvoerig en Figuur 3 gee 'n blik op die sigblad waarin evaluering gedoen is. Voorwaardelike formatering ("conditional formatting"), 'n funksie in Microsoft® Excel, word gebruik om aan te dui watter punt elke vertaling ontvang het — groen dui op 'n korrekte vertaling, geel op 'n vertaling wat met 1 gemerk is en dus nog menslike insette sal nodig hê om aanvaarbaar te wees, en rooi dui op 'n onbruikbare vertaling. Deur hierdie kleure aan te bring, kon die drie evalueerders maklik sien waar hul puntetoekenning verskil ten einde dit te bespreek en op te los.

	A	B	C	D	E	F	G	H	I	J
1	Oorspr	Dictionary Entry	Tone	POS	Noun class	English translation	Afrikaanse vertaling (Google Translate)	Afrikaanse vertaling (Bing Microsoft)	Afrikaanse vertaling (Yandex Translate)	Afrikaanse vertaling (https://english-afrikaans.co.za/#tr
2	1	phaphadzi	L	n	9	one who wanders about lookin	een wat ronddwaal op soek na iets	een wat ronddwaal om iets te soek	een wat dwaal oor op soek na iets	iemand wat ronddwaal op soek na iets
3	2	tshimbedzi	LH	n	7	dialect spoken by Vhumbedzi	dialek gepraat deur Vhumbedzi	dialek gepraat deur Vhumbedzi	dialek wat gepraat word deur Vhumbe	wat deur Vhumbedzi gepraat word,
4	3	-afhukana	H	v		be broken, shattered	stukkend wees, verpletter	gebreek, verpletter	gebreek word, verpletter	stukkend wees, verpletter
5	4	-tembisa	L	v		place in a line	plaas in 'n lyn	plek in 'n lyn	plaas in'n lyn	plek in 'n lyn,
6	5	vhuila	L	n	14	entrails	binnegoed	intrails	binnegoed	binnegoed,
7	6	mathalangwa	HHH	n	6	things scattered wide apart	dinge verspreí wyd uitmekaar	dinge wyd uitmekaar versprei	dinge gestrooi wyd uitmekaar	dinge wat wyd uitmekaar versprei is,
8	7	tshitfatfari	HHH	n	7	rough platform with railing	growwe platform met reling	rowwe platform met reling	rof platform met reling	growwe platform met reling
9	8	muhuthe	HL	n	3	species of medium-sized bird	spesies mediumgrootte voëls	spesies mediumgrootte voël	spesies van medium-grootte voël	spesies mediumgrootte voëls
10	9	mahe-a-khadzi	LLHHL	n	6	certain position of divining dic	sekere posisie van dobbelsteen	sekere posisie van goddelike dobbelste	sekere posisie van waarsêende dice	sekere posisie van waarnemende dobbelsteen
11	10	munambelo	LHL	n	3	wild plum tree	wilde pruimboom	wilde pruim boom	wilde pruim boom	wilde pruimboom
12	11	-khuba	H	v		hold fluid in mouth	hou vloeistof in die mond	hou vloeistof in die mond	hou vloeistof in die mond	hou vloeistof in die mond
13	12	-vhimba	L	v		pile up; form high towering clo	hoop op; vorm 'n hoë toring wolk	hoop op; vorm hoë toring wolk	stapel; vorm hoë toring wolk	hoop op, vorm 'n hoë toringwolk
14	13	maranga	LL	n	5	species of small cultivated cal	spesies klein gekweekte kalbas	spesies van klein gekweekte calabash	spesies van die klein gekweek kalbas	spesies klein gekweekte kalbas
15	14	muambi	L	n	1	speaker	spreker	Speaker	spreker	spreker
16	15	nnthu	H	n	9	fruit of the waterberry tree	vrugte van die waterbesboom	vrugte van die waterbessie boom	vrugte van die boom waterberry	vrugte van die waterbesboom
17	16	-potfoliya	L	v		blow a reed-flute	blaas 'n rietfluit	blaas 'n daadfluit	blaas'n riet-fluit	blaas 'n rietfluit

Figuur 3: Basisdokument vir evaluering

5.3 Verbetering van vertalings met WSpel

Een van die doelstellings van hierdie eksperiment was om die samestellingsproses sover moontlik te outomatiseer en die beskikbare hulpbronne optimaal te benut. Die navorsingspan het dus besluit om 'n speltoetser as 'n eerste stap in die redigering van die Afrikaanse vertalings in te span. Afdeling 4.3 gee besonderhede oor WSpel 15, 'n gratis Afrikaanse hulpbron. WSpel se hulpgids beskryf omvattende stappe vir installasie en gebruik met 'n paar weergawes van Microsoft® Office, maar die installasieproses is omslagtig en die beskrywing is met tye moeilik om te volg. Dit is ook net in Afrikaans beskikbaar en dus nie geskik vir leksikograwe wat nie Afrikaans magtig is nie.

Die navorsingspan het die Afrikaanse vertalings wat in die vorige stap met 'n 1 en 2 gemerk is (wat dus óf net so in die nuwe woordeboek ingesluit kan word, óf net klein veranderinge nodig het — sien Afdeling 5.2) met WSpel getoets om sodoende vas te stel of 'n speltoetser die taak van 'n redigeerder kan vergemaklik. In hierdie eksperiment het WSpel egter geen spelfoute gevind nie en kon die span dus nie die effektiwiteit daarvan as deel van die samestellingsproses behoorlik toets nie.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

5.4 *Lexonomy*-opstelling en keuse van datavelde

Lexonomy (sien Afdeling 4.6 vir 'n volledige beskrywing) is bedoel om aanpasbare oplossings vir verskillende tipes projekte vir woordeboeksamestelling te bied. Dit is dus moontlik om vooraf 'n stylblad met verpligte en opsionele velde, verskillende formaterings en teksgroottes op te stel sodat enige nuwe inskrywings daarvolgens ontwikkel word. Dit is veral nodig wanneer 'n span leksikograwe saam aan die ontwikkeling van 'n woordeboek werk en waar inskrywings een vir een bygevoeg word.

Wat hierdie sagteware egter vir ons projek uiters geskik maak, is die feit dat 'n bestaande datastel in XML-formaat opgelaai kan word en in *Lexonomy* bygewerk, verander of vir gebruik beskikbaar gestel kan word. Aangesien ons juis poog om menslike intervensie tot 'n minimum te beperk, en omdat ons reeds 'n groot deel van die voorafredigering, vertaling en verbeterings in Microsoft® Excel gedoen het, wou ons hierdie eienskap van *Lexonomy* maksimaal benut om 'n bruikbare e-woordeboek saam te stel wat later maklik bygewerk en verbeter kon word.

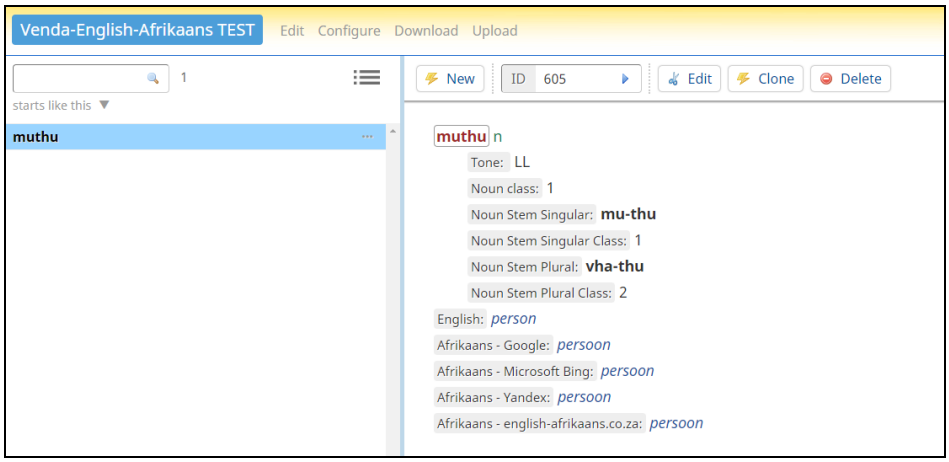
Die navorsingspan het besluit om vir die doeleindes van hierdie eksperiment te hou by die datavelde en struktuur wat in die oorspronklike woordeboek gebruik is en ook die opstelling in *Lexonomy* so te doen. Die relevante data is vervolgens uit Microsoft® Excel na 'n tekslêer oorgedra en die nodige XML-merkers is by elke kolom gevoeg. Die volgende datavelde is dus ingesluit:

— unieke ID wat aan die begin van die proses aan elkeen van die 800 terme as identifiseerder toegeken is;
— hoofwoord/lemma in Venda tesame met die klasnommer;
— Engelse vertaling; en
— Afrikaanse vertaling.

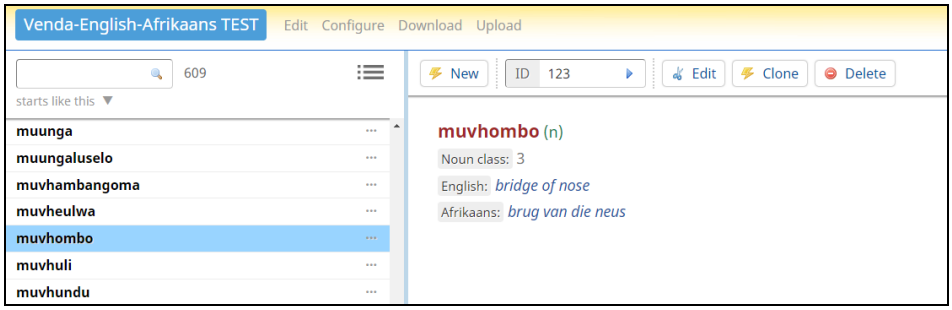
Omdat *Lexonomy* so maklik aanpasbaar is, kan addisionele datavelde soos definisies, sinonieme, gebruiksnotas, en so meer, baie maklik op 'n later stadium bygevoeg word.

Die 800 terme wat in hierdie eksperiment gebruik is, is vervolgens as 'n eerste toets van dié konsep in *Lexonomy* opgestel. Figuur 4 wys 'n skermskoot van 'n inskrywing in *Lexonomy* met al die addisionele inligting uit die Venda-woordeboek. Die span het egter besluit om vir hierdie eksperiment net die inligting wat ook in die woordeboek van Murphy (1997) opgeneem is, te vertoon om direkte vergelykings tussen die oorspronklike gedrukte weergawe en die nuwe elektroniese weergawe wat hier geskep is, te kan tref. Hierdie uiteindelike uitleg word in Figuur 5 gewys.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)



Figuur 4: Voorbeeld van 'n prototipiese inskrywing



Figuur 5: Die finale uitleg van 'n inskrywing

5.5 Uitbreiding van die lemmalys

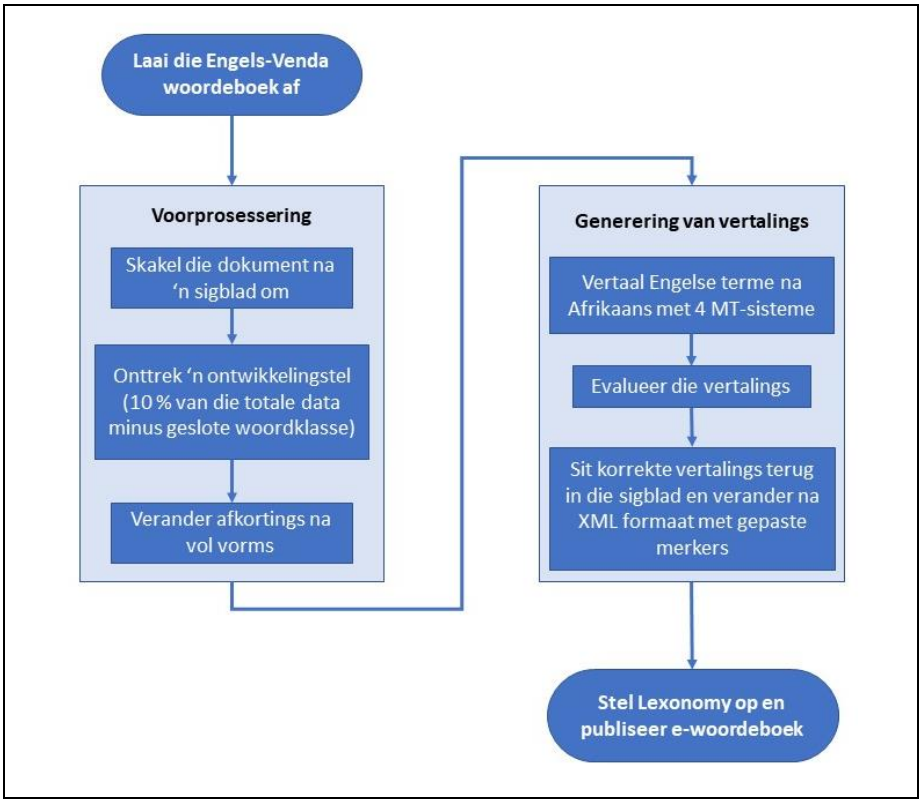
In Afdeling 3 hierbo is aangedui dat een groep teikengebruikers aanleerders van Venda as tweede of derde taal is. Dit is daarom belangrik dat lemmas met 'n hoë gebruiksfrekwensie in die lemmalys opgeneem word. As deel van die eksperiment is die Venda-korpus wat in Afdeling 4.5 hierbo beskryf is, gebruik om 'n frekwensielys op te stel. *LancsBox* is as korpusnavraagprogrammatuur gebruik. Van die 500 mees frekwente woorde wat deur *LancsBox* se frekwensie-soektog opgelewer is, kom slegs 288, d.w.s. slegs 57.6% in Murphy (1997) se woordeboek voor. Hierdie 500 woorde maak 12% van die totale korpus uit en dit is daarom belangrik dat hulle as lemmas in die woordeboek opgeneem word. Van die 27 mees frekwente woorde, wat tipies grammatiese formatiewe soos kongruensiemorfeme, ander werkwoordprefikse en naamwoordelike prefikse

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

insluit, is nie een in die woordeboek opgeneem nie. Van die mees frekwente woorde met leksikale betekenis wat nie in die woordeboek verskyn nie, is die volgende: *ṅwana* 'child' (frekwensierangorde 82), *-ḓivha* 'weet, ken' (frekwensierangorde 94), *buthano* 'byeenkoms' (frekwensierangorde 290) en *-bvela* 'uitkom, kom na' (frekwensierangorde 446). Deur die oorblywende 212 mees frekwente lemmas by die lemmalys te voeg, verhoog dit die kans op 'n suksesvolle soektog, aangesien hoë-frekwensie woorde juis dié is wat deur aanleerders nageslaan word.

Tydens die eksperiment het dit duidelik geword dat sodanige aanvulling van die lemmalys die intervensie van die leksikograaf nodig het. Daar bestaan in die eerste plek 'n hoë graad van homografie ten opsigte van die grammatiese formatiewe en die verskillende betekenisse/betekenisfunksies sal deur die leksikograaf ontrafel moet word. Tweedens sal die betekenisparafrase van die hoë-frekwensie woorde wat tot die lemmalys bygevoeg word ook deur die leksikograaf voorsien moet word.

Figuur 6 gee 'n grafiese voorstelling van die stappe in hierdie eksperiment.



Figuur 6: 'n Diagrammatiese voorstelling van die eksperiment

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

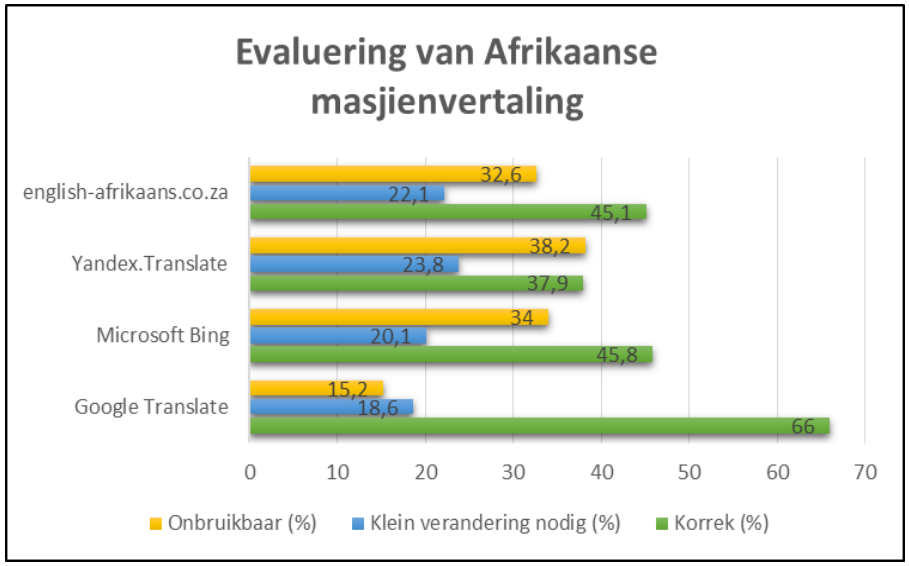
6. Evaluering

6.1 Masjienvertaling

Uit die handmatige evaluering van die masjienvertaling van Engels na Afrikaans blyk dit duidelik dat die vier sisteme nie almal ewe effektief is nie. Die tabel (Tabel 1) onder gee enkele algemene indrukke wat in Afdeling 6.2 met voorbeelde aangevul word.

	Google Translate	Microsoft® Bing	Yandex. Translate	english-afrikaans.co.za
Korrekte vertaling	528	366	303	361
Klein verandering nodig	149	161	190	177
Onbruikbare vertaling	122	272	306	261
Onseker	1	1	1	1
Totaal	**800**	**800**	**800**	**800**

Tabel 1: Samevatting van die handmatige evaluering



Figuur 7: Grafiese voorstelling van die samevatting van die handmatige evaluering

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Google Translate lewer dus die meeste vertalings wat direk in 'n woordeboek ingesluit kan word sonder dat enige menslike intervensie nodig is, met 528 (66 %) van die inskrywings wat as korrek gemerk is. *Yandex.Translate* blyk die meeste naredigering te verg omdat 190 (23.8 %) van die inskrywings klein veranderinge nodig het en 306 (38.2 %) heeltemal onbruikbaar is en dus deur 'n linguis voorsien sal moet word.

Dit is ook interessant dat 613 van die 800 gevalle (77 %) ten minste een korrekte vertaling bevat en dat al vier vertalers in 168 gevalle (21 %) 'n korrekte (maar nie noodwendig dieselfde) vertaling lewer en dus 'n hoë sekerheidsgraad behaal. Die voordeel van 'n elektroniese woordeboek is natuurlik dat die samestellers nie hier 'n keuse vir net een vertaling hoef te maak nie en selfs al vier kan insluit omdat die formaat nie soveel beperkings op die artikellengte stel nie.

In 26 gevalle is ten minste een van die vertalings met 'n 1 gemerk, wat beteken dat net klein veranderings deur 'n leksikograaf of taalkenner nodig sal wees om 'n aanvaarbare vertaling te lewer wat ook in die woordeboek gevoeg kan word.

Die volgende afdeling bespreek van die vertaaluitdagings in meer besonderhede.

6.2 Vertaaluitdagings

6.2.1 (Semi-)outomatiese korreksies

Sekere uitdagings kan (semi-)outomaties reggestel word, bv. woordorde, ortografie (koppeltekens, hoofletters, spasiëring) en vertalings van meervoud versus enkelvoud.

Woordorde

Die woordorde in baie van die Afrikaanse vertalings volg die woordorde van die Engelse weergawe, bv. *become wealthy* word in vertaalsisteem D as *word ryk* vertaal, terwyl vertaalsisteme A en B die woordvolgorde korrek as *ryk word* vertaal. Nog 'n voorbeeld is: *bring nearer* wat deur A en B korrek as *nader bring* vertaal word, terwyl C en D die volgorde omruil *bring nader.* Hierdie woordvolgordeprobleem duik slegs by werkwoorde op, in gevalle waar die Engelse werkwoord uit meer as een woord bestaan. By nadere beskouing het dit duidelik geword dat in die meerderheid van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie, woordorde die probleem is (30 gevalle). Vir Afrikaanstalige gebruikers van die woordeboek behoort dit geen probleem te wees nie.

Ortografie

Daar is heelwat probleme rakende ortografiese aangeleenthede soos bv.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

die gebruik van koppeltekens al dan nie, en ook die vas- en losskryf van woorde. Enkele voorbeelde is: *akasia-boom* vs. *akasiaboom*, *Venda-taal* vs. *Vendataal* en *swart mamba slang* vs. *swart mamba-slang*. Dié ortografiese probleem kom veral in die geval van eiename soos boom- en plantname voor, maar kan maklik met behulp van 'n Afrikaanse speltoetser opgelos word. Daar is bevind dat slegs 16 van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie, met koppeltekens en die vas- en losskryf van woorde te doen het.

In sommige gevalle kom 'n vertaalsisteem met 'n hoofletter vorendag, bv. *gaste* vs. *Gaste*. 'n Semi-outomatiese vind-en-vervang proses is 'n maklike oplossing, maar daar moet versigtig te werk gegaan word sodat eiename nie outomaties na kleinletters verander word nie.

Spasiëring is deurgaans 'n probleem in vertaalsisteem C wanneer dit kom by 'n spasie tussen die lidwoord *'n* en die voorafgaande woord, bv. *braai oor 'n oop vuur, opgesluit in 'n klein hok.* In dié gevalle is 'n outomatiese vind-en-vervang prosedure die aangewese oplossing.

Vertalings van meervoud vs. enkelvoud

Die Engelse woord *species* kom slegs in die meervoud voor, maar die korrekte weergawe in Afrikaans is die enkelvoud *spesie*, bv. die korrekte vertaling van *species of medium-sized bird* sou wees *spesie mediumgrootte voël*. Die vier vertaalsisteme is nie konsekwent met die enkelvoud- en meervoud-vertalings nie. 'n Semi-outomatiese vind-en-vervang prosedure is 'n maklike oplossing van die probleem. Dieselfde geld vir 'n voorbeeld soos *fruit* in *fruit of muhukhuma*. Volgens die *Macmillan English Dictionary for Advanced Learners* (2002: 571) is die meervoud van *fruit* óf *fruit* óf *fruits* en dit is waarskynlik die rede waarom die vertaalsisteme inkonsekwent is met die Afrikaanse vertaling. Na aanleiding van die Venda enkelvoud-naamwoordklas 9 van *muhukhuma* sou die korrekte vertaling dus *vrug* en nie *vrugte* wees nie. Weereens is 'n semi-outomatiese vind-en-vervang prosedure is 'n maklike oplossing. Dit is slegs van toepassing op 11 van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie.

Hoflikheidsvorm "u"

In sommige Afrikaanse vertalings word die hoflikheidsvorm "u" gebruik, bv. in *inciting others to do something that causes discord, while staying uninvolved oneself* wat deur sisteme A en D vertaal word as *ander aan te spoor om iets te doen wat onenigheid veroorsaak, terwyl u onbetrokke bly*. Sulke gevalle kan ook met 'n semi-outomatiese vind-en-vervang prosedure benader word.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

6.2.2 Menslike intervensie

Menslike intervensie is nodig in gevalle waar konteks (polisemie), kulturele begrippe en domeinspesifieke terme 'n uitdaging is. Korreksies moet in sulke gevalle handmatig aangebring word.

Konteks

Daar is verskeie gevalle van polisemie opgemerk waar 'n woord verskeie betekenisse in verskillende kontekste het, soos in die voorbeeld:

> *bar or pole for barring cattle kraal gate > kroeg of paal vir die versperring van beeskraalhek > kroeg of paal vir die versperring van beeste kraal hek > bar of paal behalwe vir beeste kraal hek bar of paal vir blokkeer beeskraal hek.*

In 'n ander konteks sou *bar* met *kroeg* vertaal kon word, maar in hierdie spesifieke konteks is die betekenis *paal* die aangewese een in die konteks van *beeskraalhek.*

Dieselfde geld vir die polisemiese Engelse werkwoorde *pull, draw* wat in sisteme A en B as *trek, teken* vertaal word. In die gegewe konteks is slegs *trek* korrek. 'n Soortgelyke voorbeeld is *rim* en *edge* wat albei as *rand* in Afrikaans vertaal word. Die Engelse weergawe *make a rim or edge on a basket* word deur al die sisteme verkeerdelik as *maak 'n rand of rand op 'n mandjie* vertaal.

Kultuurgebonde begrippe

Een van die Engelse definisies van die Venda-werkwoordstam *-xa* word gegee as *lose all counters in mufuvha game*. Agtergrondkennis (soos gevind by https://www.bead.game/games/traditional/mefuvha) is nodig om die korrekte vertaling in Afrikaans te identifiseer, naamlik *tellers* en nie *toonbanke* nie. Vertaalsisteme B en C gee wel die korrekte vertaling.

Die naamwoord *chief* in die volgende voorbeeld is ook kultuurgebonde: *hut of chief's uncle, brother, son.* Die korrekte vertaling is dus *hut van hoofman se oom, broer, seun.* Sisteme B en D gee die korrekte vertaling as *hoofman*, terwyl sisteme A en C onderskeidelik *owerste* en *hoof* as vertaling gee wat wel in 'n ander konteks korrek sou wees.

Hoewel kultuurgebonde begrippe hoogswaarskynlik 'n baie lae frekwensie in Venda-tekste het, het hulle tog kultuurhistoriese waarde.

Domeinspesifieke terme

Leemtes in die leksikons van die vertaalsisteme kom na vore in die geval van domeinspesifieke terme, soos bv. plant- en voëlname, waar dikwels 'n letterlike vertaling gedoen word. Die term *Cape robin* word deur drie van die vier vertaalsisteme as *Kaapse robin* vertaal terwyl een sisteem slegs die Engelse term weergee. Die korrekte Afrikaanse term is *janfrederik.* 'n Voor-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

beeld van 'n plantnaam is *cabbage tree* wat deur al vier sisteme letterlik vertaal word as *koolboom* of *kool boom* in plaas van *kiepersol*.

Meerdere vertaalekwivalente

Dit is interessant om op te merk dat volgens 'n woordeboek soos Pharos se *Verklarende Afrikaanse woordeboek* woorde soos *os* en *bees* as ekwivalent beskou word, en ook dat *blom* as ekwivalent vir *blossom* gegee word. Dit is derhalwe nie vreemd dat die vertaalsisteme daarmee akkoord gaan nie.

Afwesigheid van korrekte vertaalekwivalente

Indien daar geen korrekte vertaalekwivalent bestaan nie, sal die leksikograaf uiteraard 'n vertaalekwivalent moet verskaf, soos in die geval van die werkwoord *khakhamedza* met die Engelse vertaling *take aback*. Al vier Afrikaanse vertalings wat verskaf word, is foutief, naamlik *skrik, neem terug* en *neem uit die veld geslaan* (laasgenoemde word deur twee van die vertaalsisteme verskaf). Die korrekte vertaling sou wees *verstom* of *uit die veld slaan*.

7. Gevolgtrekking

7.1 Gevolgtrekkings en samevatting

Die bevindinge van ons aanvanklike ondersoek na die beskikbaarheid van meertalige woordeboeke vir Afrikatale met skaars hulpbronne, soos vir Venda, het daarop gedui dat die enigste Venda–Engels–Afrikaans woordeboek al vir 'n geruime tyd uit druk is. Die behoefte aan so 'n woordeboek het intussen ontstaan as gevolg van twee tipes gebruikers, naamlik direkte en indirekte teikengebruikers. Direkte teikengebruikers sluit taalaanleerders soos onderwysstudente in, terwyl indirekte teikengebruikers daarna streef om die woordeboekdata te gebruik om die taal tegnologies te ontwikkel, onder andere vir doeleindes van masjienvertaling.

In hierdie artikel word die optimale benutting van gratis elektroniese/aanlyn hulpbronne vir die saamstel van 'n bruikbare drietalige e-woordeboek vir Venda, Engels en Afrikaans wat mettertyd maklik bygewerk kan word, ondersoek. Die benadering wat gevolg is, behels 'n eksperiment waarin die samestellingsproses so ver moontlik geoutomatiseer is om besparing in terme van tyd en mens-ure teweeg te bring. Engels word as 'n brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Die gratis beskikbare hulpbronne wat gebruik is, sluit in 'n Venda–Engels woordeboek, vier Engels–Afrikaans masjienvertaalsisteme, 'n Afrikaanse speltoetser, korpusondersoekprogramme en 'n program vir woordeboeksamesteling. Hierdie eksperiment is op 10% ewekansig uitgesoekte woordeboekinskrywings gebaseer wat vier woordsoortklasse insluit, naamlik naamwoorde, werkwoorde,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

bywoorde en adjektiewe. Geslote woordklasse is uitgesluit. Die handmatige evaluering deur eerstetaalsprekers is ook op hierdie data uitgevoer.

Die algemene bevindinge van die eksperiment is dat daar — soos te wagte — sekere beperkings op so 'n semi-outomatiese proses met gratis hulpbronne is, wat 'n sekere mate van menslike intervensie verg. Hoewel die saamgestelde e-woordeboek nie as 'n finale produk beskou kan word nie, bied die wolkgebaseerde, oopbron woordeboekskrywer en -publiseerder *Lexonomy* die geleentheid vir menslike insette soos deur byvoorbeeld leksikograwe, linguiste, ens. om die nodige bywerkings op 'n gebruikersvriendelike wyse te doen. Dit is te danke aan die aanpasbaarheid en maklike uitleg van *Lexonomy*. Verdere woordklasse, veral geslote woordklasse, kan met behulp van 'n grammatika soos Poulos (1990) se *A Linguistic Analysis of Venda* vergelyk word, en deur menslike intervensie aangevul word, bv. die verskillende tipes voornaamwoorde. Dit is interessant om op te merk dat 'n geslote woordklas soos voegwoorde goed verteenwoordig is in die Venda–Engels Woordeboek (Murphy 1997), naamlik 27 voegwoorde altesaam, terwyl Poulos (1990) slegs die 15 mees frekwente voegwoorde insluit.

'n Beduidende voordeel van die (semi-)outomatiese proses wat in hierdie artikel beskryf is, is die besparing aan mens-ure wat benut is in vergelyking met tyd wat normaalweg deur leksikograwe spandeer word aan die ontwikkeling van formele woordeboeke. Die konsepvoorstel wat geformuleer is, is nuttig vir die daarstel van meertalige aanlyn woordeboeke, saamgestel met behulp van gratis beskikbare aanlyn of elektroniese hulpbronne. Ongelukkig is selfs die beskikbaarstelling van elektroniese hulpmiddels, veral waar kleiner tale soos Afrikaans en Venda ter sprake is, ook nie altyd so volhoubaar nie. Een van die masjienvertaalsisteme wat ons in Julie 2021 vir die eksperiment gebruik het, english-afrikaans.co.za, was byvoorbeeld teen Januarie 2022 nie meer aanlyn beskikbaar nie. Webdienste soos hierdie se betroubaarheid word dan in twyfel getrek.

Die resulterende drietalige woordeboek wat saamgestel is, is reeds aanlyn as 'n *Lexonomy*-woordeboek beskikbaar om as 'n bewys van die konsep waarop verdere werk kan bou te dien[8]. Die feit dat die databasis wat die woordeboek onderlê in 'n masjienleesbare formaat, naamlik XML, afgelaai kan word, is belangrik vir indirekte teikengebruikers vir hergebruik om elektroniese hulpbronne te ontwikkel, veral vir hulpbronarm tale. Aangesien die navorsingspan ten gunste is van maksimale toegang tot elektroniese hulpbronne, veral vir die Afrikatale, stel ons ook die XML-weergawe van die woordeboek vir navorsers en ander gebruikers gratis beskikbaar op die webblad van SADiLaR (2022), 'n organisasie wat die skep van digitale hulpbronne vir die tale van Suid-Afrika ten doel het.

7.2 Toekomstige werk

'n Formele evaluering van die mate waarin die Afrikaanse vertalings akkurate

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

vertaalekwivalente van die Venda lemmas weergee, sal veel bydra tot die waarde van die eksperiment. Hiervoor is die insette van moedertaalsprekers van Venda nodig. Dit sal verder ook interessant wees om die lemmalys van die woordeboek aan leksikografiese meetinstrumente soos dié van Prinsloo en De Schryver (2002) te toets ten einde vas te stel tot watter mate die verskillende alfabetiese strekke toereikend behandel is.

Die eksperiment soos hierbo beskryf maak dit moontlik om dié werkswyse na ander tale uit te brei en so binne 'n relatiewe kort tydperk en met heelwat minder mens-ure elektroniese woordeboeke bestaande uit verskillende taalpare beskikbaar te stel. Ten opsigte van die Afrikatale bestaan daar byvoorbeeld geen woordeboeke waarin beide die brontaal en die teikentaal Afrikatale is nie. Die metodologie soos hierbo beskryf maak die saamstel van sulke woordeboeke 'n haalbare onderneming.

Die woordeboek self kan verder uitgebrei word deur die datavelde uit te brei. Velde kan byvoorbeeld geskep word vir gebruiksvoorbeelde wat uit die korpus onttrek word en hoë-frekwensie kollokasies. Hierdie prosesse kan semi-outomaties met behulp van gratis korpusnavraagprogrammatuur uitgevoer word. Benewens die uitbreiding van die lemmalys soos deur frekwensie bepaal, is 'n kritiese evaluering van die huidige lemmalys nodig, met inagneming van die teikengebruiker. Verouderde of argaïese lemmas hoort waarskynlik nie in 'n aanleerderswoordeboek tuis nie en so 'n evaluering sal in samewerking met 'n Venda-spesialis gedoen moet word.

Eindnotas

1. Sien https://www.w3.org/standards/xml/core vir 'n volledige beskrywing van hierdie formaat.
2. http://autshumato.sourceforge.net/
3. https://en.wikipedia.org/wiki/Google_Translate
4. https://en.wikipedia.org/wiki/Microsoft_Translator
5. https://en.wikipedia.org/wiki/Yandex.Translate
6. https://wspel.wordpress.com/
7. https://wspel.wordpress.com/
8. 'n Gratis profiel kan op die Lexonomy-platform geregistreer word by https://www.lexonomy.eu/ en daarna is die konsepweergawe van die woordeboek by https://www.lexonomy.eu/POCVenEngAfr/ te sien.

Erkennings

— **CBOLD**

Die outeurs is dank en erkenning verskuldig aan die samestellers van die *Comparative Bantu OnLine Dictionary* (http://www.cbold.ish-lyon.cnrs.fr/) wat hulle woordeboekdata in elektroniese formaat beskikbaar stel vir nie-kommersiële gebruik deur ander navorsers.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

— **SADiLaR**

Hierdie projek is moontlik gemaak met ondersteuning van SADiLaR (2022), 'n navorsingsinfrastruktuur wat deur die Departement van Wetenskap en Tegnologie van die Suid-Afrikaanse regering gestig is as deel van die Suid-Afrikaanse navorsingsinfrastruktuur-padkaart (SARIR).

— **Nasionale Navorsingstigting (NNS)**

Hierdie navorsing is finansiëel ondersteun deur die NNS. Die toekenninghouers (Unieke verwysings: S E Bosch (Toekenning nr. 109384) en E Taljard (Toekenning nr. 77735)) bevestig dat die menings, bevindinge en gevolgtrekkings of aanbevelings wat in enige NNS-ondersteunde navorsing uitgespreek word, hul eie is en dat die toekenningsinstansie geen aanspreeklikheid in dié verband aanvaar nie.

Bibliografie

Woordeboeke

Labuschagne, F.J. en L.C. Eksteen. 2010. *Pharos verklarende Afrikaanse woordeboek.* Kaapstad: Pharos Woordeboeke. Beskikbaar: https://www.pharosaanlyn.co.za/tuis

Murphy, M.L. 1997. *Venda: CBOLD (Comparative Bantu OnLine Dictionary).* Beskikbaar: http://www.cbold.ish-lyon.cnrs.fr/

Rundell, M. (Red.). 2002. *Macmillan English Dictionary for Advanced Learners*. Second edition. Oxford: Macmillan Education.

Van Warmelo, N.J. 1937. *Tshivenḓa–English Dictionary*. Pretoria: Staatsdrukker.

Wentzel, P.J. en T.W. Muloiwa. 1976. *Drietalige Elementêre Woordeboek / Trilingual Elementary Dictionary: Venda–Afrikaans–English.* Pretoria: Universiteit van Suid-Afrika.

Wentzel, P.J. en T.W. Muloiwa. 1982. *Ṭhalusamaipfi ya nyambotharu yo khwiniswaho: Luvenḓa–Luvhuru–Luisimane / Verbeterde drietalige woordeboek: Venḓa–Afrikaans–Engels / Improved Trilingual Dictionary: Venḓa–Afrikaans–English.* Pretoria: Universiteit van Suid-Afrika.

Ander bronne

Atkins, B.T.S., J. Clear en N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.

Bartolomé-Díaz, B. en F. Frontini. 2020. Building a Domain-specific Bilingual Lexicon Resource with *Sketch Engine* and *Lexonomy*: Taking Ownership of the Issues. *Proceedings of the 2020 Globalex Workshop on Linked Lexicography, May 2020, Marseille, France*: 62-68. Marseille: European Language Resources Association. Beskikbaar: https://aclanthology.org/2020.globalex-1.11.pdf

CBOLD. 1997–2003. *Comparative Bantu OnLine Dictionary*. Beskikbaar: http://www.cbold.ish-lyon.cnrs.fr/.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Departement van Hoër Onderwys en Opleiding. 2015. National Qualifications Framework Act (67/2008): Revised Policy on the Minimum Requirements for Teacher Education Qualifications. *Government Gazette/Staatskoerant*, 19 Februarie 2015. Beskikbaar: https://bit.ly/31xp8rV

Eckart, T., S. Bosch, D. Goldhahn, U. Quasthoff en B. Klimek. 2019. Translation-based Dictionary Alignment for Under-resourced Bantu Languages. Eskevich, Maria, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek en Milan Dojchinovski (Reds.). 2019. *2nd Conference on Language, Data and Knowledge (LDK 2019):* 17:1-17:11. Schloss Dagstuhl — Leibniz-Zentrum für Informatik: Dagstuhl Publishing. Beskikbaar: http://drops.dagstuhl.de/opus/volltexte/2019/10381/pdf/OASIcs-LDK-2019-17.pdf

European Lexicographic Infrastructure (ELEXIS). 2020. *Opening up Dictionaries, Linguistic Data and Language Tools for European Communities.* [Brochure]. Beskikbaar: https://elex.is/wp-content/uploads/2019/03/Print_Publicity_Brochure.pdf

Groves, M. en K. Mundt. 2015. Friend or Foe? Google Translate in Language for Academic Purposes. *English for Specific Purposes* 37: 112-121.

Jakubíček, M., M. Měchura, V. Kovář en P. Rychlý. 2018. Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. 2018. *The XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, Slovenia, July 17–21, 2018*. 65-67. Beskikbaar: https://euralex2018.cjvt.si/wp-content/uploads/sites/19/2020/08/Euralex2018_book_of_abstracts_FINAL.pdf

Jurafsky, D. en J.H. Martin. 2009. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey: Prentice Hall.

McKellar, C.A. en H.J. Groenewald. 2012. Frequency-based Data Selection for Statistical Machine Translation with Scarce Resources. Ndinga-Koumba-Binza, H.S. en S.E. Bosch (Reds.). 2012. *Language Science and Language Technology in Africa: A Festschrift for Justus C. Roux*: 271-290. Stellenbosch: SUN MeDIA.

Měchura, M. 2017. Introducing Lexonomy: An Open-source Dictionary Writing and Publishing System. Kosem, I. et al. (Hrsg.). 2017. *Electronic Lexicography in the 21st Century, Proceedings of eLex 2017 Conference, 19–21 September 2017, Leiden, the Netherlands:* 662-679. Brno: Lexical Computing CZ s.r.o.

Moors, C., I. Wilken, K. Calteaux en T. Gumede. 2018. Human Language Technology Audit 2018: Analysing the Development Trends in Resource Availability in all South African Languages. *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists SAICSIT '18: Technology for Change, 26–28 September 2018, Port Elizabeth, South Africa:* 296-304. New York: The Association for Computing Machinery (ACM). Beskikbaar: https://doi.org/10.1145/3278681.3278716

N̂emuṭamvuni, M.E. 2018. *Investigating the Effectiveness of Available Tools for Translating into Tshivenḓa*. M.A.-tesis, Universiteit van Suid-Afrika, Pretoria. Beskikbaar: http://hdl.handle.net/10500/25563

Poulos, G. 1990. *A Linguistic Analysis of Venda*. Pretoria: Via Afrika.

Prinsloo, D.J. en G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, Anna and Claus Povlsen (Reds.). 2002. *Proceedings of the Tenth EURALEX*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002: 483-494. Copenhagen: Center for Sprogteknologi, Københavns Universitet.

SADiLaR. 2022. *South African Centre for Digital Language Resources.* Beskikbaar: https://sadilar.org/index.php/en/

Sager, J.C. 1994. *Language Engineering and Translation: Consequences of Automation.* Amsterdam/ Philadelphia: John Benjamins.

Stemle, E.W., A. Abel en V. Lyding. 2019. Language Varieties Meet One-Click Dictionary. Kosem, I. et al. (Reds.). 2019. Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 *Conference, 1–3 October* 2019, *Sintra, Portugal*: 537-546. Brno, Czech Republic: Lexical Computing CZ s.r.o. Beskikbaar: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_31.pdf