http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

Die verifiëring, verfyning en toepassing van leksikografiese liniale vir Afrikaans

D.J. Prinsloo, *Departement Afrikatale, Universiteit van Pretoria, Pretoria, Suid-Afrika (danie.prinsloo@up.ac.za)*

Opsomming: Leksikografiese liniale vir Afrikaans en die Afrikatale is 'n dekade oud en word algemeen gebruik in die samestelling van woordeboeke. Die samestellers het dit tot dusver nie nodig geag om hierdie liniale te verifieer of te verfyn nie. Kritiek is egter uitgespreek op die samestelling van die Afrikaanse Liniaal en dit word in hierdie artikel opgevolg deur 'n poging tot verifiëring van die bestaande liniaal asook 'n herberekening of verfyning van die Afrikaanse Liniaal. Vir die verifiëring word 'n sogenaamde *stresfaktor* as die basiese benadering gebruik. Dertien liniale word bereken deur middel van 13 subkorpusse van Afrikaans wat doelbewus só saamgestel is dat dit die mees *ongunstige* toestande vir die bestaande liniaal skep ten einde te bepaal tot watter mate sodanige liniale afwyk van die Afrikaanse Liniaal. Verfyning van die Afrikaanse Liniaal word gedoen deur die gemiddelde van vyf liniale te neem wat gebaseer is op dié tipes tekskategorieë wat internasionaal in sogenaamde gebalanseerde en verteenwoordigende korpusontwerpe voorgehou word. Vir dié doel word korpusse saamgestel van koeranttekste, kreatiewe skryfwerk, religieuse tekste en sowel formele as informele taalgebruik. Die gemiddelde waardes van die Afrikaanse Liniaal en die verfynde liniaal word dan as 'n nuwe sogenaamde 2010 Liniaal vir Afrikaans voorgehou. Ten slotte word die onlangs voltooide dele XII en XIII van die *Woordeboek van die Afrikaanse Taal* kortliks met die Afrikaanse Liniaal gemeet.

Sleutelwoorde: LEKSIKOGRAFIE, WOORDEBOEKE, ALFABETIESE KATEGORIEË, LEKSIKOGRAFIESE LINIAAL, BALANS, KORPUS, KORPUSDATA, OORBEWERKING, ONDERBEWERKING, KORPUSONTWERP

Abstract: The Verification, Refinement and Application of Lexicographic Rulers for Afrikaans. Lexicographic rulers for Afrikaans and the African languages are a decade in existence and are generally used for the compilation of dictionaries. To date the compilers of these rulers did not feel the need to verify or to refine these rulers. Criticism has however been expressed against the ruler for Afrikaans and this is followed up in this article by an effort to verify the existing ruler and a recalculation or refinement of the Afrikaans Ruler. For verification a so-called *stress factor* is used as the basic approach. Thirteen rulers are calculated by means of 13 sub-corpora of Afrikaans purposely compiled to create the most *unfavourable* circumstances for the existing ruler in order to ascertain to what extent these rulers deviate from the Afrikaans Ruler. Refinement of the Afrikaans Ruler is done by calculating the average of five rulers based on those categories that are internationally being used in so-called balanced and representative corpora. For this purpose, corpora are compiled from newspaper texts, creative writing, religious texts and formal as well as informal texts. The average values of the Afrikaans Ruler and the refined ruler are then presented as a new so-called 2010 Ruler for Afrikaans. Finally the recently completed volumes XII and XIII of the *Woordeboek van die Afrikaanse Taal* will briefly be compared with the Afrikaans Ruler.

Lexikos 20 (AFRILEX-reeks/series 20: 2010): 390-409

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

Keywords: LEXICOGRAPHY, DICTIONARIES, ALPHABETIC CATEGORIES, LEXICOGRAPHIC RULER, BALANCE, CORPUS, CORPUS DATA, OVER-TREATMENT, UNDER-TREATMENT, CORPUS DESIGN

Inleiding

Aan die begin van die nuwe millennium het Prinsloo en De Schryver sogenaamde *leksikografiese liniale* vir die Afrikatale en vir Afrikaans en Engels ontwerp in reaksie op die talle inkonsekwenthede op makrostrukturele vlak wat betref die balans ten opsigte van alfabetiese kategorieë wat hulle in woordeboeke van dié tale teëgekom het. Leksikografiese liniale is vir die eerste keer in 2002 by die Tiende Konferensie van die European Association for Lexicography (EURALEX) in Kopenhagen formeel aan die internasionale gemeenskap bekendgestel (Prinsloo en De Schryver 2002). Die praktiese gebruik van die liniaal vir Afrikaans is uitvoerig gedemonstreer in Prinsloo en De Schryver (2003). Verdere noemenswaardige geleenthede was die gebruik van die liniaal as vertrekpunt vir verdere evaluering van die WAT (De Schryver 2005) en die publikasie van liniale vir al 11 die amptelike landstale van Suid-Afrika (Prinsloo en De Schryver 2005).

Hierdie liniale is sedertdien aan die Suid-Afrikaanse Nasionale Leksikografie-eenhede gedemonstreer en by die opleiding van leksikograwe geïnkorporeer. Leksikografiese liniale word ook deur talle vryskutwoordeboekmakers gebruik en is ook geïntegreer in die woordeboekprogram *Tshwanelex*, 'n gesofistikeerde rekenaarprogram vir die samestelling van woordeboeke.

In die eerste aantal jare sedert 2000 was die aandag beperk tot die uitwerk van die liniale en die gebruik daarvan in die samestelling van nuwe woordeboeke of vir bywerking/hersiening en kritiese analise van bestaande woordeboeke. Geen aandag is gegee aan die moontlike verfyning van die liniale soos wat meer korpusdata deur die jare beskikbaar geword het nie. Die akkuraatheid van die liniale is aanvanklik ook nie bevraagteken nie. Die ontwerpers het dit derhalwe nie nodig geag om die akkuraatheid van die liniale te verifieer of om hulle te probeer verfyn nie. Kritiek is vir die eerste keer uitgespreek, en wel teen die Afrikaanse Liniaal in Botha (2005) in reaksie op De Schryver (2005) se stellings ten opsigte van oor- en onderbewerking in die WAT.

> I do not believe that the inclusion of the desk dictionaries in the ruler is warranted, owing to their inherent deficiencies. I therefore have some doubt whether the data resources on which the ruler is based, can be considered as balanced and can give frequency counts that accurately reflect Afrikaans. (Botha 2005: 78)

Hierdie kritiek teen die Afrikaanse Liniaal het daartoe gelei dat die samestelling van die Afrikaanse Liniaal in die afgelope vyf jaar opnuut onder die loep geneem is en die resultaat in hierdie artikel aangebied word.

Die kernvraag wat gevra moet word vir die verfyning van die liniale is of

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

verandering van die databasis wat vir die berekening van die liniaal gebruik word, sê byvoorbeeld verskillende, onverwante korpusse, verskillende liniale sal oplewer, en as dit die geval is tot watter mate die liniale sal verskil: ingrypend/totaal verskillend of bloot in 'n geringe mate?

Die doel van hierdie artikel is dus om die liniaal vir Afrikaans te verifieer en te verfyn. Verifiëring geskied met behulp van 'n aantal onverwante liniale wat uit diverse korpusse van Afrikaans saamgestel is en die berekening van 'n verfynde liniaal vir Afrikaans geskied op basis van vyf liniale, elk verteenwoordigend van die kategorieë koerantberigte, kreatiewe skryfkuns, religieuse tekste, formele en informele en gesproke taalgebruik. Die seleksie van hierdie kategorieë is gebaseer op die ontwerpe van die *Brown Corpus of Standard American English* (BROWN) en die *Lancaster-Oslo/Bergen Corpus (*LOB*)*, die *Longman/Lancaster English Language Corpus* en die ICE. Ten slotte word die resente dele XII (P) en XIII (Q–R) van die WAT ooreenkomstig die Afrikaanse Liniaal gemeet. Ten einde die Afrikaanse liniale en in besonder die begrippe verwante en onverwante liniale, in perspektief te stel, word die 11 liniale vir die amptelike Suid-Afrikaanse landstale as vertrekpunt geneem. Die eerste stap is om die 11 liniale voor te stel en hulle met mekaar te vergelyk met spesifieke verwysing na die grootste verskille in die omvang van die alfabetiese kategorieë in die onderskeie tale.

Motivering vir en die ontwerp van leksikografiese liniale

Die besluit om leksikografiese liniale te ontwerp, spruit uit waarnemings deur De Schryver en Prinsloo van oënskynlike inkonsekwenthede in lemmaseleksie en ongebalanseerde bewerking van lemmas in woordeboeke, veral ten opsigte van oor- of onderbewerking van alfabetiese kategorieë. Tipiese gevalle is dié waar die leksikograaf die samestelling van die woordeboek oorentoesiasties aanpak en dan stoom verloor wat oorbewerking van die eerste paar alfabetiese kategorieë en onderbewerking van die laaste kategorieë tot gevolg het. Vergelyk Kriel (1983) as 'n tipiese voorbeeld in dié verband ten opsigte van die omvang van die bewerkings in die kategorie A versus T in (1).

(1) Pukuntšu

aka, *a.ka.* (*-ile, -etše*), lieg, leuens vertel, jok, onwaarheid spreek (dial. kyk: *aketša*).
aka, *a.ka*, inhaak, vashaak, haak, aanhaak, soen, omarm, lieg, liefkoos; *akwa*, gehaak/ingehaak word; *akêla*, haak vir; *akelana*, mekaar liefkoos, vriendskaplik verkeer; *akelwa*, ingehaak word vir; *akiwa*, ingehaak word; *ake, ga, sa*, nie (in)haak nie; *akê*, mag/moet haak of inhaak; *moaki*, haker; *baaki*, hakers.
akalala, *a ka la.la*, sweef, hang oor, oorhang; *akalalêla*, sweef vir/oor; *akalatša*, laat sweef, vlerke oopsprei om te sweef, *akaladitše*, het laat sweef; *se bone nong go-. go wa fase ke ga lona*, hoogmoed kom tot 'n val; *akalatšwa*, genoodsaak om te sweef; *akalalwa*; gesweef word; *akalêla*, hang/sweef oor, wydsbeen staan oor; *akaletše*; het gesweef oor; *moakaladi*, persoon wat sweef.
akama, *a ka.ma*, verwonder/verbaas wees; *akamela*, inlaat (bemoei) met; *akametša*, (laat) verbaas, verbasing wek, aangaap, toeroep; akametšwa, verbaas/aangegaap word, toegeroep word.

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

akere, *'a kê.'rê*, akker.
aketša, *a ke.tša*, leuen vertel, lieg, jok; *akeditše*, het (gelieg) 'n leuen vertel; *sa aketše*, nie lieg nie.
akga, *a.kaga*, werp, gooi, slinger, swaai, beweeg; *akgaakga*, heen en weer beweeg (soos branders), slinger, skommel; *akgaakgwa*, heen en weer geslinger word; - *diatla*, arms swaai, met leë hande loop; - *dinao*, voet in die wind slaan; *akgwa*, beweeg,/geslinger word; - *akgêga*, skommel, swaai; -*akgêla*, slinger, swaai, werp; *akgêla*, slinger na/vir, tou om die horings gooi, met 'n vangtou vang, uitkrap, soos kole uit 'n vuur; *akgelwa*, geslinger word, gevang word met 'n tou; - *dikobo*, klere uitpluk.

tsirikana, *'tsi'ri ka.na*, klink.
tsirima, *'tsi'ri.ma*, klink, lui, uitspuit, vorentoe spring.
tsirimetša, 'tsi'ri me.tša, laat klink, vasbyt, laat lui, styf vasbind.
tsirinya, *'tsi'ri.nya*, laat klink, lui.
tširoga, *'tši ro.ga*, wakker skrik, senuweeagtig word, opskrik, moedeloos word.
tširogo *'tši ro.gô*, impuls.
tširoša *'tši ro.ša*, wek, skrikmaak.

Die teenoorgestelde (Atkins: mondelinge mededeling, Maart 2004) kom ook voor waar die leksikograaf 'stadig' begin en dan oorentoesiasties raak soos wat die projek vorder of dit kan geskied as gevolg van beleids- of bestuursveranderinge in die samestelling van 'n woordeboek wat oor 'n lang tydperk saamgestel word.

Leksikografiese liniale is ontwerp om die leksikograaf te help om 'n goeie balans te handhaaf ten opsigte van die relatiewe grootte van die alfabetiese kategorieë ten opsigte van die aantal bladsye en die aantal lemmas wat per kategorie bewerk word. Liniale vergestalt daardie *inherente balans* tussen die alfabetiese kategorieë van elke taal en beantwoord die eenvoudige vraag van hoe groot moet kategorie A in die woordeboek wees in verhouding tot kategorie B, kategorie C, ens. In eenvoudige Afrikaans geformuleer, is die vraag bloot "hoe weet die leksikograaf wanneer kategorie A genoegsaam bewerk is en dit tyd is om aan te skuif na kategorie B". Sodanige balans is veral noodsaaklik in gevalle waar 'n voorafbepaalde maksimum aantal bladsye wat die woordeboek mag beslaan deur die uitgewer gestel is. Hierdie aspek van woordeboekmaak is uiters belangrik — vele individuele samestellers en selfs groot woordeboeke het al deur die jare in die slaggat getrap van oor- of onderbewerking van sekere alfabetiese kategorieë in hulle woordeboeke. Svensén, alhoewel slegs met verwysing na die bestudering van woordeboeke as basis vir so 'n balans, stel die beginsel nietemin onomwonde:

> A decision must also be made as to what fraction of the whole dictionary each initial letter may occupy, so that the size of the finished dictionary can be kept under control during the course of the project. The percentages for each of the various initial letters in a given entry language are fairly constant, and, if such calculations have not already been done by others, it is wise to examine the distribution in a number of dictionaries. (Svensén 1993: 242)

Thorndike (Landau 2001: 360-362) ontwerp 'n sogenaamde bloksisteem vir die distribusie van woordeboekinskrywings ten opsigte van eerste letters. Hy ver-

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

deel die alfabet in 105 blokke waarin ongeveer dieselfde gewig aan elke blok toegeken word met die doel om 'n ewewigtige verspreiding van leksikale eenhede deur die alfabet te reflekteer. Thorndike se blokstelsel vir die distribusie van woordeboekinskrywings ken byvoorbeeld vier blokke vir E en 13 blokke aan S toe. Vergelyk Tabel 1 wat 'n uittreksel vir E en S uit die Thorndike-sisteem is.

Tabel 1: Die blokke E en S in Thorndike se bloksisteem vir die distribusie van woordeboekinskrywings per eerste letter (Landau 2001: 361)

...	**S-81** sau–sd	**S-88** splo–stas
E-29 e–elk	**S-82** sea–seo	**S-89** stat–stov
E-30 ell–en	**S-83** sep–shio	**S-90** stow–sucg
E-31 eo–exb	**S-84** ship–sinf	**S-91** such–swar
E-32 exc–ez	**S-85** sing–smd	**S-92** swas–sz
...	**S-86** sme–sors	...
S-80 s–sat	**S-87** sort–spln	

Landau (2001: 360) merk tereg op:

> If one's word list shows that E has as many entries as S, for example, one should suspect that whoever selected the terms for E was far more permissive than the selector for S, and adjust the word list accordingly.

Absolute versus relatiewe liniaalwaardes

Die doel van Prinsloo en De Schryver met die samestelling van leksikografiese liniale was die daarstelling van 'n meetinstrument vir die grootte van alfabetiese kategorieë as 'n *basiese riglyn*. Alhoewel berekenings per kategorie tot een desimaal afgerond word, suggereer dit nie 'n normeringswaarde dat A 10.0% groot behoort te wees en dat die geringste afwyking verkeerd of ontoelaatbaar is nie. Wat dit wel suggereer, is dat enige substansiële afwyking, sê byvoorbeeld groter as 2%, die moontlikheid van oor- of onderbewerking impliseer en dat dit raadsaam sal wees om dan volgens Landau (2001: 360) die 'woordelys te verstel'. 'n Oorbewerking van 1% van 'n groot kategorie waarvan die liniaalriglyn 10% is, verteenwoordig 'n absolute oorbewerking van 1% en 'n relatiewe oorbewerking van ongeveer 10%, maar 'n oorbewerking van 1% op 'n liniaalriglyn van 1% is 'n oorbewerking van 100%. (Vergelyk Prinsloo en De Schryver (2003: 110) vir afsonderlike berekenings van die relatiewe en absolute waardes van die Afrikaanse Liniaal.) As riglyn in die praktiese samestelling van 'n woordeboek is die absolute waarde van meer belang, dit wil sê om te probeer om nie meer as 'n persentasiepunt of twee van die liniaalwaarde af te wyk nie.

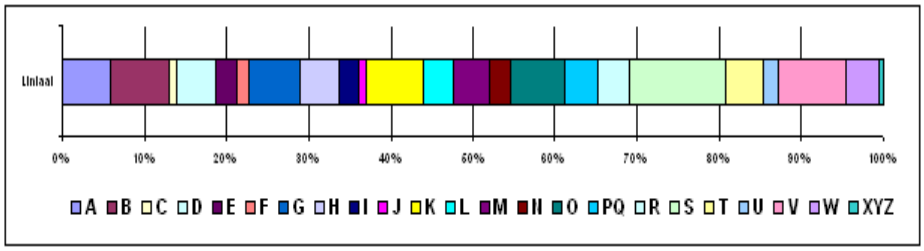
Die aanvanklike liniaal vir Afrikaans is gebaseer op twee tipes bronne, woordeboeke en 'n korpus vir Afrikaans, maar gegewe die kritiek van Botha (2005) en die feit dat daar tans baie meer korpusmateriaal beskikbaar is as 'n

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

dekade gelede, word slegs korpusmateriaal vir die verifiëring en verfyning van die Afrikaanse liniaal in hierdie studie gebruik.

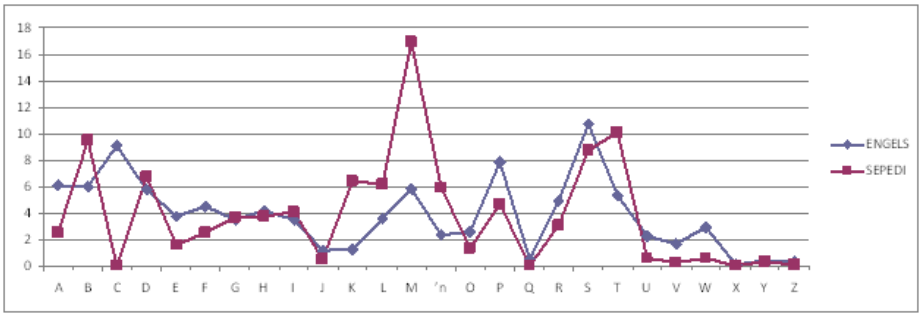
Liniale vir Afrikaans en die ander amptelike landstale van Suid-Afrika

Prinsloo en De Schryver (2005) gee die volledige stel soos in Bylaag A en die liniaal vir Afrikaans soos in Figuur 1. Prinsloo en De Schryver (2002: 488) bevind dat liniaalberekenings op ongelemmatiseerde en gelemmatiseerde korpusdata dieselfde resultate lewer ('n korrelasiekoëffisiëntwaarde r = 0.991 (met r = 1.0 as die perfekte korrelasiewaarde, nl. twee identiese getallereekse)). Berekenings is gebaseer op ongelemmatiseerde tipes ('types', verskillende woorde) wat in die korpus voorkom en korpusgrootte word ooreenkomstig tekens ('tokens', die aantal woorde in die korpus) aangegee.



Figuur 1: Liniaal vir Afrikaans in % [**P** en **Q**; **X**, **Y** en **Z** saamgevoeg] (Prinsloo en De Schryver 2005)

Dit wat 'n gesonde balans vir die grootte van 'n spesifieke alfabetiese kategorie vir taal A verteenwoordig, kan presies die teenoorgestelde vir taal B beteken. So byvoorbeeld is die kategorie **M** in Sepedi baie groot (17%) in vergelyking met Engels (6%), terwyl die kategorie **C** in Engels weer baie groot is (9%) in vergelyking met Sepedi (0%), ens. Vergelyk die liniale vir Engels en Sepedi in Figuur 2 as 'n voorbeeld van *onverwante* liniale.

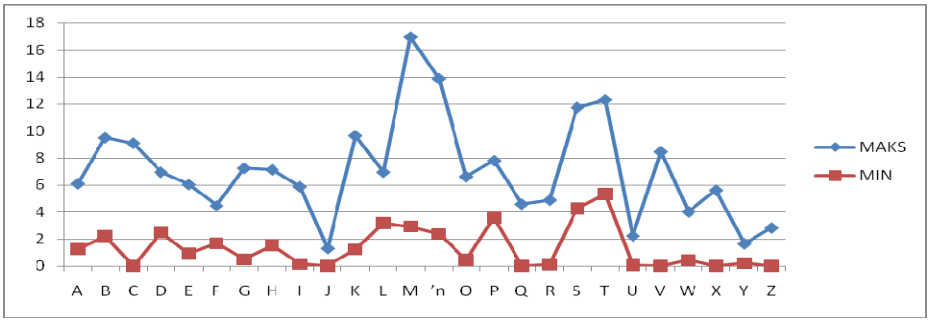


Figuur 2: 'n Vergelyking van die liniale vir Engels en Sepedi

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

Die belangrikste basiese beginsel is dat elke taal se liniaal uniek is. Geeneen van die 11 landstale se inherente balans in alfabetiese kategorieë kom ooreen met dié van 'n ander taal nie, selfs nie eers vir nouverwante tale soos Sepedi, Setswana en Sesotho nie (vergelyk Prinsloo 2006).

Die minimum en maksimum waardes per alfabetiese kategorie vir al die landstale word aangedui in Figuur 3.



Figuur 3: Minimum en maksimum liniaalwaardes vir alfabetiese kategorieë van die 11 amptelike landstale

Verifiëring van die Afrikaanse Liniaal

Die blote gemiddelde of die opstapeling van meer liniale wat op willekeurige seleksie van woordeboeke en subkorpusse gebaseer is, sou in beginsel kon bydra tot die verifiëring en selfs tot die verfyning van die Afrikaanse Liniaal. Wat verifiëring betref, is daar besluit om 'n sogenaamde *stresfaktor* as die basiese benadering te gebruik deur korpusse doelbewus só saam te stel dat dit die mees *ongunstige* toestande vir die bestaande Afrikaanse Liniaal skep en dan te bepaal of dit 'n reeks onverwante Afrikaanse liniale tot gevolg het. Vir hierdie doel is 13 subkorpusse saamgestel en 13 afsonderlike liniale bereken. Tabel 2 dui die aard, grootte en samestelling van die subkorpusse aan. 'n Kriptiese omskrywing van hulle negatiewe aard asook hulle korrelasie met die Afrikaanse Liniaal word in Tabel 3 gegee.

Tabel 2: Samestelling van die 13 subkorpusse vir die verifiëring van die Afrikaanse Liniaal

No.	Aard van / tipe teks	Korpusgrootte in aantal woorde (tekens)	Aantal verskillende woorde (tipes)	Bron
1	Literêre werk	105 008	10 267	*Versamelde werke* C.J. Langenhoven (1933 en 1934)

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

2	Koerantmateriaal: *Burger* en *Beeld*, woorde wat meer as een keer voorkom	141 513 937	496 135	Seleksie uit die *Pharos*-toetskorpus (*Media24*-argief)
3	Koerantmateriaal: *Burger* en *Beeld*, woorde wat slegs een keer voorkom	141 513 937	499 551	Seleksie uit die *Pharos*-toetskorpus (*Media24*-argief)
4	Borduurwerk	64 968	4 609	*Borduursteke vir Suid-Afrika* (Eaton 1989)
5	Laslappie- en appliekwerk	28 131	4 353	*Suid-Afrikaanse boek van laslappie- en appliekwerk* (Turpin-Delport 1988)
6	Tuinbou	49 146	6 758	*Suid-Afrikaanse tuin* (Gilbert 1985)
7	E-postekste	4 346	1 172	E-posbus — eie versameling
8	Religieuse tekste	919 002	14 986	*Afrikaanse Bybel*
9	Pornografie	10 225	2 062	*Loslyf* (2000)
10	Landbou	8 620 710	177 886	Seleksie uit *Landbouweekblad* (*Media24*-argief)
11	Koerantmateriaal: *Rapport*	5 000 829	98 290	Seleksie uit *Rapport* (*Media24*-argief)
12	Akademiese taal	40 661	4 134	*UP Strategiese Plan* (2002–2005)
13	Gesproke taalgebruik	2 007	574	*Kyknet: Robinson Regstreeks*. April 2010

Tabel 3: Die negatiewe aard en korrelasie van die 13 subkorpusse vir die verifiëring van die Afrikaanse Liniaal

No.	**Subliniaal-projeknaam**	**Negatiewe kriteria**	**Korrelasie-koëffisiënt**
			r =
1	Langenhoven	Verouderde Afrikaans, klein korpus	0.972333
2	Pharos > 1	Domeinspesifiek, hoë(r) frekwensie, onnatuurlike benadering deur die korpus in twee onverwante dele te verdeel op grond van frekwensie	0.939373

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

3	Pharos = 1	Domeinspesifiek, baie lae en onnatuurlike frekwensieseleksie	0.955235
4	Borduur	Baie spesifieke onderwerp, herhalende tipiese woordeskat, baie klein korpus	0.948399
5	Laslap	Baie spesifieke onderwerp, herhalende tipiese woordeskat, baie klein korpus	0.959935
6	Tuin	Baie spesifieke onderwerp, herhalende tipiese woordeskat, baie klein korpus	0.943563
7	E-pos	Baie informele teks, baie klein korpus	0.927709
8	Bybel	Domeinspesifiek	0.947419
9	Pornografie	Baie spesifieke onderwerp, herhalende tipiese woordeskat, baie klein korpus, lae register, baie informele teks	0.942511
10	Landbou	Baie domeinspesifiek	0.968453
11	*Rapport*	Domeinspesifiek	0.963554
12	*UP Strat. Plan*	Baie formeel, hoë register	0.928534
13	Gesproke taal	Baie klein korpus, gesproke taal, informeel	0.916724
		Gemiddeld	0.947211

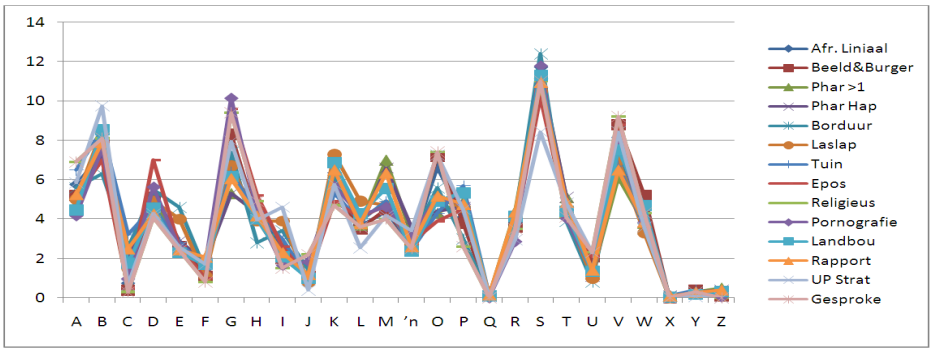
'n Analise van die verskillende tipes taalgebruik in Tabel 3 dui op besondere verskille. In die geval van die verouderde tekste in 1 is die gebruik van verouderde woordeskat en spelwyses soos *begint* (begin), *had* (het ... gehad), *vammelewe* (vroeër/lank gelede), *posiesie* (posisie), *poliesie* (polisie), *ergens* (êrens), *tamelik* (redelik), *taggentig* (tagtig), *seg* (sê) opmerklik. Ten opsigte van uiteenlopende onderwerpe is dit ook te verwagte dat 'n liniaal slegs gebaseer op laslappiewerk waarskynlik onverwant sal wees aan 'n liniaal saamgestel vir 'n tuinboukorpus of dat die balans in woordeskat tussen die Bybel en 'n pornografietydskrif verskillend sal wees. Die tipiese en herhalende taalgebruik in byvoorbeeld die pornografietekste sentreer rondom 'n klein aantal vulgêre woorde. *Hapax legomena* (woorde wat slegs een keer in 'n korpus voorkom) word in baie studies geïgnoreer as irrelevant vir taalkundige gevolgtrekkings. Korpuslinguiste staan ook skepties ten opsigte van rigtinggewende gevolgtrekkings wat op baie klein korpusse gemaak word soos byvoorbeeld die gesproketaalkorpus van 2 000 woorde en 'n liniaal wat op slegs 574 verskillende woorde gebaseer is.

In al die gevalle sou 'n mens verwag dat dié unieke (negatiewe) eienskappe die liniaalwaardes sal versteur.

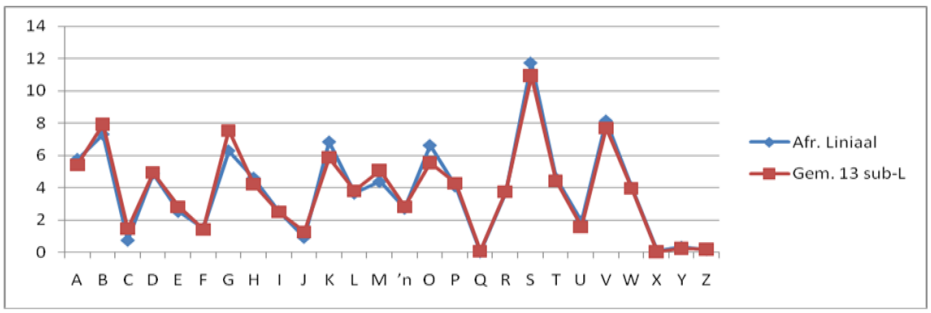
Dit is egter merkwaardig dat die liniale van elkeen van hierdie 13 uiteenlopende 'onvriendelike' subkorpusse 'n korrelasiekoëffisiënt van hoër as 0.9% met die Afrikaanse Liniaal vertoon en die gemiddelde van hierdie liniale so hoog as 0.95 is soos grafies in Figuur 4 voorgestel word.

Selfs uit hierdie sogenaamde onvriendelike liniale blyk die inherente balans ten opsigte van alfabetiese kategorieë in Afrikaans duidelik en Figuur 5 dui dié noue korrelasie van die Afrikaanse Liniaal met die gemiddelde van die 13 subliniale aan.

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149



Figuur 4: Afrikaanse Liniaal versus 13 onvriendelike Afrikaanse subliniale



Figuur 5: Die Afrikaanse Liniaal versus die gemiddelde van die 13 onvriendelike Afrikaanse subliniale

'n Verfynde liniaal vir Afrikaans

Dit is voor die handliggend dat 'n verfynde liniaal ten beste uit 'n gebalanseerde en verteenwoordigende korpus van Afrikaans ontwikkel moet word, veral in die lig daarvan dat veel meer Afrikaanse teks tans in elektroniese formaat beskikbaar is as 10 jaar gelede.

Daar bestaan egter nie 'n waterdigte gebalanseerde of verteenwoordigende korpusontwerp vir enige taal nie. Korpuslinguiste kon tot dusver nie eens eenstemmigheid bereik oor die presiese betekenis in korpusverband van die terme 'gebalanseerd' en 'verteenwoordigend' nie. Bekende korpusontwerpe soos die Brown/LOB-, *Longman/Lancaster Oslo Bergen-* en die ICE-korpusse in Tabel 4 is bloot pogings om soveel moontlik tipiese taalgebruik in die ontwerp en in die fisiese korpus te inkorporeer. Hierdie debat lê egter buite die bestek van hierdie artikel (sien Biber 1993, Summers 1993, Kilgarriff 1997, Kennedy 1998, Kruyt en Dutilh 1997, Otlogetswe 2007, en Atkins en Rundell 2008 vir uitvoerige bespreking).

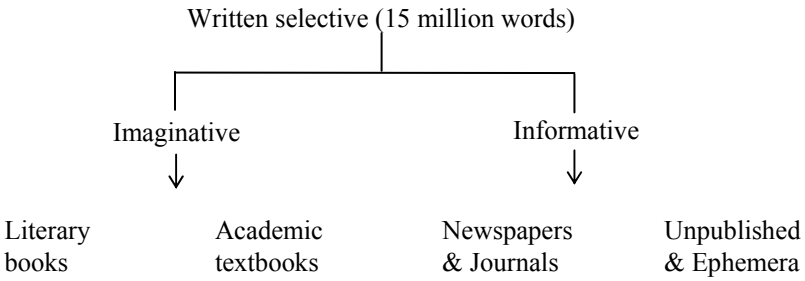
http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

Tabel 4: Brown/LOB, *Longman/Lancaster Oslo Bergen* en die ICE

(1) Brown/LOB-korpusontwerp

PRESS: REPORTAGE (44 texts)	FICTION: GENERAL (29 texts)
PRESS: EDITORIAL (27 texts)	FICTION: MYSTERY (24 texts)
PRESS: REVIEWS (17 texts)	FICTION: SCIENCE (6 texts)
RELIGION (17 texts)	FICTION: ADVENTURE (29 texts)
SKILLS AND HOBBIES (36 texts)	FICTION: ROMANCE (29 texts)
POPULAR LORE (48 texts)	HUMOR (9 texts)
BELLES-LETTRES (75 texts)	MISCELLANEOUS: GOVERNMENT
LEARNED (80 texts)	& HOUSE ORGANS (30 texts)

(2) 'n Seleksie van die *Longman/Lancaster English Language Corpus*



(3) Die ICE-korpusontwerp vir geskrewe tekste

Written Texts (200)	Non-printed (50)	**Non-professional writing (20)**	untimed student essays (10) student examination scripts (10)
		Correspondence (30)	social letters (15) business letters (15)
	Printed (150)	**Academic writing (40)**	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Non-academic writing (40)	humanities (10) social sciences (10) natural sciences (10) technology (10)
		Reportage (20)	press news reports (20)
		Instructional writing (20)	administrative/regulatory (10) skills/hobbies (10)
		Persuasive writing (10)	press editorials (10)
		Creative writing (20)	novels/stories (20)

Vir die samestelling van 'n verfynde liniaal word 'n meer simplistiese ontwerp voorgehou wat gebaseer is op die ontwerpe van die Brown/LOB-, *Longman/ Lancaster Oslo Bergen-*, en die ICE-korpusse. Die verfyning geskied op basis van vyf liniale, elk verteenwoordigend van die kategorieë koerantberigte, kreatiewe

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

skryfkuns, religieuse tekste, formele dokumente en informele en gesproke taalgebruik.

Tabel 5: Korpusontwerp vir die berekening van die verfynde Afrikaanse Liniaal

Subkorpus	**Samestelling**	**Woorde**	**Verskillende woorde**
Koerantberigte	Koerante: *Beeld* en *Burger*	142 013 488	995 686
Kreatiewe skryfkuns	Digbundels, kortverhale	4 283 294	165 599
Religieuse tekste	Preke, Bybeltekste	975 361	16 490
Formele dokumente	Wette, regeringsdokumente	125 320	7 147
Informele en gesproke taal	E-pos, stokperdjies, geselstaal, gesproke taal	198 225	16 950

Dieselfde gewig word aan elke kategorie toegeken ongeag die grootte van die subkorpus deur die gemiddeld van die vyf aparte liniale te bereken, vergelyk Tabel 6.

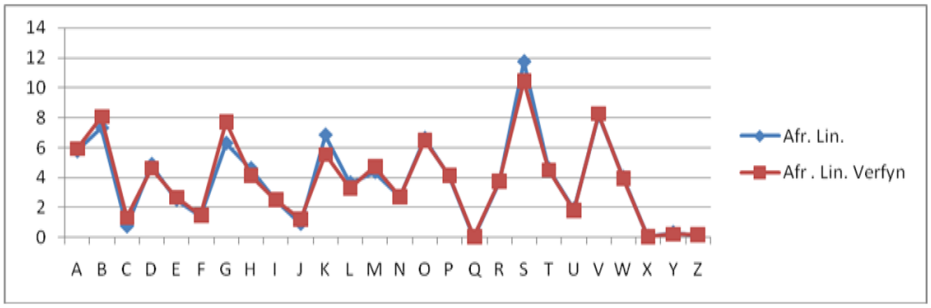
Tabel 6: Die vyf sublinale, die Afrikaanse Liniaal en die Verfynde Liniaal

	Koerant	Kreatief	Religieus	Formeel	Informeel	Afr. Liniaal	Afr. Lin. Verfyn
A	4.8	5.6	6.8	6.5	6	5.8	5.9
B	7.5	8.3	7.9	8.7	8	7.3	8.1
C	2.3	1	0.3	2.1	0.9	0.7	1.3
D	5.2	5	4.1	4.5	4.3	4.9	4.6
E	2.7	3	2.5	3	2.3	2.5	2.7
F	1.9	1.6	0.8	1.6	1.6	1.5	1.5
G	5.3	8.2	9.2	7	8.9	6.3	7.7
H	4.3	4.1	4.8	3.4	4.1	4.6	4.2
I	2.1	2.7	1.7	3.8	2.4	2.5	2.5
J	1.2	1.2	2.1	0.7	0.9	0.9	1.2
K	6.5	6	4.8	4.4	6	6.8	5.6
L	3.7	3.3	3.6	2.5	3.4	3.7	3.3
M	6.8	4.6	4.0	4.1	4.2	4.4	4.7
N	3.4	2.7	2.5	3	2.1	2.7	2.7
O	4.4	6.2	7.4	7.4	7.1	6.6	6.5
P	4.8	4.2	2.7	4.9	4.2	4.1	4.2
Q	0.1	0.1	0.0	0.1	0	0.0	0.1
R	4.2	3.5	3.4	4.2	3.3	3.7	3.7

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

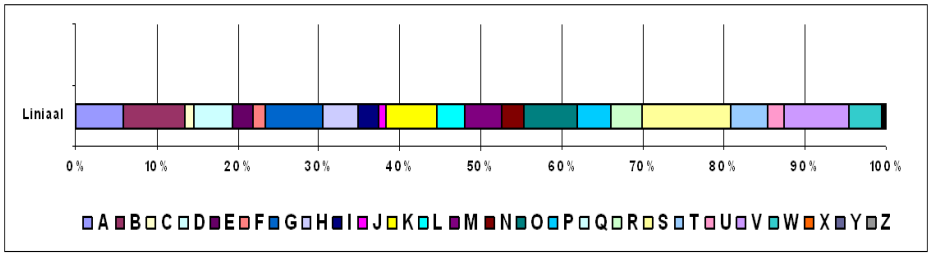
S	11.4	10.1	11.0	9.1	10.6	11.7	10.5
T	5.1	4.2	4.2	4.4	4.7	4.6	4.5
U	1.1	1.6	2.3	2	2	1.9	1.8
V	6.5	8.6	9.2	8.5	8.6	8.1	8.3
W	3.9	3.9	4.4	3.7	3.8	4.0	3.9
X	0.1	0	0.0	0.1	0	0.1	0
Y	0.3	0.3	0.3	0.1	0.2	0.3	0.2
Z	0.4	0.2	0.0	0.1	0.2	0.2	0.2

Die korrelasiekoëffisiëntwaarde van die Afrikaanse Liniaal en die verfynde liniaal is $r = 0.98$ en dié noue verwantskap word in Figuur 6 grafies aangedui.



Figuur 6: 'n Verfynde liniaal vir Afrikaans

Die gemiddelde van die Afrikaanse Liniaal en die Verfynde Liniaal word in Figuur 7 as die Afrikaanse Liniaal 2010 voorgehou.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
5.9	7.7	1.0	4.8	2.6	1.5	7.0	4.4	2.5	1.1	6.2	3.5	4.6	2.7	6.6	4.1	0.04	3.7	11.1	4.6	1.9	8.2	4.0	0.1	0.3	0.2

Figuur 7: Die saamgestelde 2010-Liniaal vir Afrikaans

Leksikografiese liniale kan benewens 'n persentasiewaarde per letter van die alfabet, ook vir praktiese doeleindes in enige aantal dele uitgedruk word. Prinsloo en De Schryver (2003: 123) verdeel die Afrikaanse liniaal in 179 blok-

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

ke. In Tabel 7 word die liniaal vir formele dokumente (ook Tabel 6 kolom 5) in 100 dele opgebreek wat elk dus een persent van die liniaal verteenwoordig.

Tabel 7: 'n Bloksisteem vir Afrikaans bestaande uit 100 dele

1	AANR	21	DOKU	41	JARE	61	OPLE	81	SYFE
2	ABSO	22	EENH	42	KATA	62	ORDE	82	TEKO
3	AFGE	23	EKSP	43	KLER	63	PAPI	83	TEWE
4	AKKR	24	ERKE	44	KONT	64	PERM	84	TOER
5	ANDE	25	FAKS	45	KUND	65	POGI	85	TRUS
6	ARTS	26	FORM	46	LAND	66	PRIM	86	UITG
7	BEDO	27	GEBO	47	LENI	67	PROM	87	UITV
8	BEHO	28	GEHA	48	LOSI	68	RADI	88	VARI
9	BENA	29	GEMA	49	MANL	69	REFE	89	VERB
10	BESI	30	GEPA	50	MEGA	70	REGU	90	VERK
11	BEST	31	GESO	51	MINI	71	RESO	91	VERS
12	BEWA	32	GEWE	52	MPUM	72	RUST	92	VERT
13	BLYW	33	GRON	53	NASP	73	SATE	93	VISS
14	BREE	34	HAND	54	NETW	74	SELF	94	VOLW
15	BYLA	35	HERO	55	NOTA	75	SIMU	95	VOOR
16	CHOR	36	HOOF	56	OMGE	76	SKUL	96	VYFD
17	CV	37	IDEN	57	ONDE	77	SONS	97	WATE
18	DEBA	38	INFR	58	ONTE	78	STAA	98	WERK
19	DESE	39	INRI	59	OORB	79	STER	99	WILD
20	DINO	40	INTE	60	OORV	80	STUD	100	ZIMB

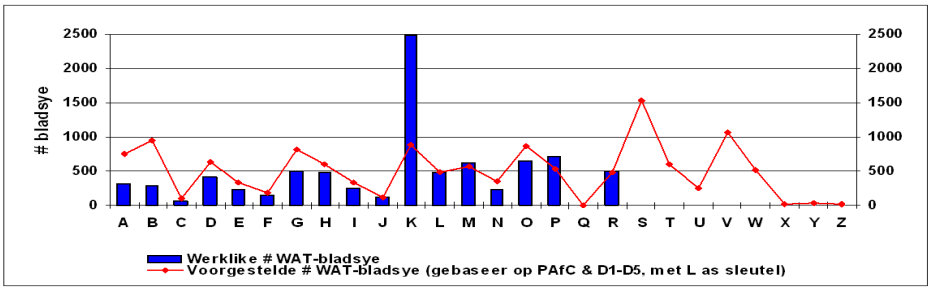
Die bloksisteem kan met vrug gebruik word by onder meer die bestuur van 'n woordeboekprojek. Daar kan te eniger tyd bepaal word of die projek op skedule is, byvoorbeeld ingevolge tyd en die aantal toegelate bladsye. Indien die woordeboek byvoorbeeld binne twee jaar voltooi moet word en nie meer as 1 000 bladsye mag beslaan nie, beteken dit dat een blok per week afgehandel moet word en die totale lengte nie 10 bladsye per blok mag oorskry nie. Die aantal lemmas per blok en die gemiddelde lengte van die artikels kan vooraf bepaal word. Werkverrigting deur die onderskeie leksikograwe en selfs die vergoeding aan deeltydse samestellers kan op dié manier gemeet word. Dit is wat Prinsloo en De Schryver (2003) met die term 'effektiewe vordering' bedoel.

Dele XII en XIII van die WAT gemeet aan die Afrikaanse Liniaal

Vir hierdie doel word die oorspronklike WAT-liniaal gebruik wat in Prinsloo en De Schryver (2003) vir die evaluering van die WAT ontwerp is, ten einde te bepaal tot watter mate die alfabetiese kategorieë P, Q en R met die voorspelde

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

liniaalwaardes korreleer. Figuur 8 is die bygewerkte grafiek wat ook die P-, Q- en R-waardes reflekteer.



Figuur 8: Die bygewerkte WAT-liniaal

Vir P is die liniaalwaarde 534 bladsye en die werklike aantal bladsye 718, wat ooreenkomstig die liniaal 'n matige oorbewerking suggereer. Vir Q is die liniaalwaarde 2.7 en die werklike bladsye 3 en vir R 486 en 507 respektiewelik. Kategorieë Q en R korreleer dus presies met die liniaal. Voorspelde liniaalwaardes ooreenkomstig bladsye vir die onvoltooide dele S tot W volgens die oorspronklike berekening (Prinsloo en De Schryver 2003) is S = 1 529.2, T = 598.8, U = 248.7, V = 1 061.3 en W = 523.7.

Slotopmerkings

'n Voortreflike leksikograaf moet deurgaans stry teen alle vorme van inkonsekwentheid en wanbalans tydens die samestelling van 'n woordeboek. Die oor- of onderbewerking van alfabetiese kategorieë in woordeboeke is nie bloot 'n 'tegniese' of 'akademiese' aangeleentheid nie — dit het direkte implikasies vir die woordeboekgebruiker wanneer die inligtingsaanbod verskil byvoorbeeld van te veel tot te min in dieselfde woordeboek. Leksikografiese liniale maak 'n bydrae tot die kwaliteit van woordeboeke en die gebruikersperspektief. Dit wil ook voorkom of hierdie balans maklik bepaalbaar is deurdat selfs baie klein korpusse van so min as 'n paar honderd tekens, soos in die geval van e-pos- en geselstaalkorpusse, reeds voldoende is om dié balans aan te dui.

Literatuurlys

Afrikaanse Bybel. http://www.bybel.co.za

Atkins, B.T. Sue en M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Biber, D. 1993. Using Register-diversified Corpora for General Language Studies. *Computational Linguistics* 19(2): 219-241.

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

Botha, W. 2005. Concurrent Over- and Under-treatment in Dictionaries. A Response. *International Journal of Lexicography* 18(1): 77-87.

Brown Corpus of Standard American English. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html.

De Schryver, G.-M. 2005. Concurrent Over- and Under-treatment in Dictionaries — The *Woordeboek van die Afrikaanse Taal* as a Case in Point. *International Journal of Lexicography* 18(1): 47-75.

Eaton, J. 1989. *Borduursteke vir Suid-Afrika. 'n Volledige gids*. Kaapstad: Delos.

Gilbert, Z. 1985. *Die Suid-Afrikaanse tuin. Maand vir maand.* Tweede uitgawe. Johannesburg: Central News Agency.

ICE. http://www.ucl.ac.uk

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Londen/Nieu-York: Longman.

Kilgarriff, A. 1997. Putting Frequencies in the Dictionary. *International Journal of Lexicography* 10(2): 135-155.

Kriel, T.J. 1983. *Pukuntšu Noord-Sotho–Afrikaans, Afrikaans–Noord-Sotho Woordeboek*. Pretoria: J.L. van Schaik.

Kruyt, J.G. en M.W.F. Dutilh. 1997. A 38 Million Words Dutch Text Corpus and its Users. *Lexikos* 7: 229-244.

Kyknet: *Robinson Regstreeks*. DSTV. http://www.dstv.com

Lancaster-Oslo/Bergen Corpus. http://www.essex.ac.uk/linguistics/clmt/w3c/corpus_ling/content/corpora/list/private/LOB/lob.html.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Tweede uitgawe. New York/Cambridge: Cambridge University Press.

Landbouweekblad: Media 24.

Langenhoven, C.J. 1933. *Versamelde werke. Deel V*. Kaapstad: Nasionale Pers.

Langenhoven, C.J. 1934. *Versamelde werke. Deel VI.* Kaapstad: Nasionale Pers.

Loslyf. Afrikaanse sekstydskrif. Oktober 2000. 6(5).

Media24. http://www.media24.co.za.

Otlogetswe, T.J. 2007. *Corpus Design for Setswana Lexicography*. Ongepubliseerde Ph.D.-proefskrif. Pretoria: Universiteit van Pretoria.

Pharos-toetskorpus. http://www.pharos.co.za

Prinsloo, D.J. 2006. Compiling a Bidirectional Dictionary Bridging English and the Sotho Languages: A Viability Study. *Lexikos* 16: 193-204.

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, A. en C. Povlsen (Reds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Kopenhagen: Sentrum vir Taaltegnologie, Universiteit van Kopenhagen.

Prinsloo, D.J. and G.-M. de Schryver. 2003. Effektiewe vordering met die *Woordeboek van die Afrikaanse Taal* soos gemeet in terme van 'n multidimensionele Liniaal. Botha, W. (Red.). 2003. *'n Man wat beur. Huldigingsbundel vir Dirk van Schalkwyk*: 106-126. Stellenbosch: Buro van die WAT.

Prinsloo D.J. en G.-M. de Schryver. 2005. Managing Eleven Parallel Corpora and the Extraction of Data in all Official South African Languages. Daelemans, W., T. du Plessis, C. Snyman en L. Teck (Reds.). 2005. *Multilingualism and Electronic Language Management. Proceedings of the 4th*

http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

International MIDP Colloquium, 22–23 September 2003, Bloemfontein, South Africa: 100-122. Studies in Language Policy in South Africa 4. Pretoria: Van Schaik.

Rapport: Media 24.

Svensén, B. 1993. *Practical Lexicography: Principles and Methods of Dictionary-Making*. Oxford: Oxford University Press.

Summers, Della. 1993. Longman/Lancaster English Language Corpus — Criteria and Design. *International Journal of Lexicography* 6(3): 181-208.

Tshwanelex. http://tshwanedje.com/tshwanelex/

Turpin-Delport, L. 1988. *Die Suid-Afrikaanse boek van laslappie- en appliekwerk*. Kaapstad: C. Struik.

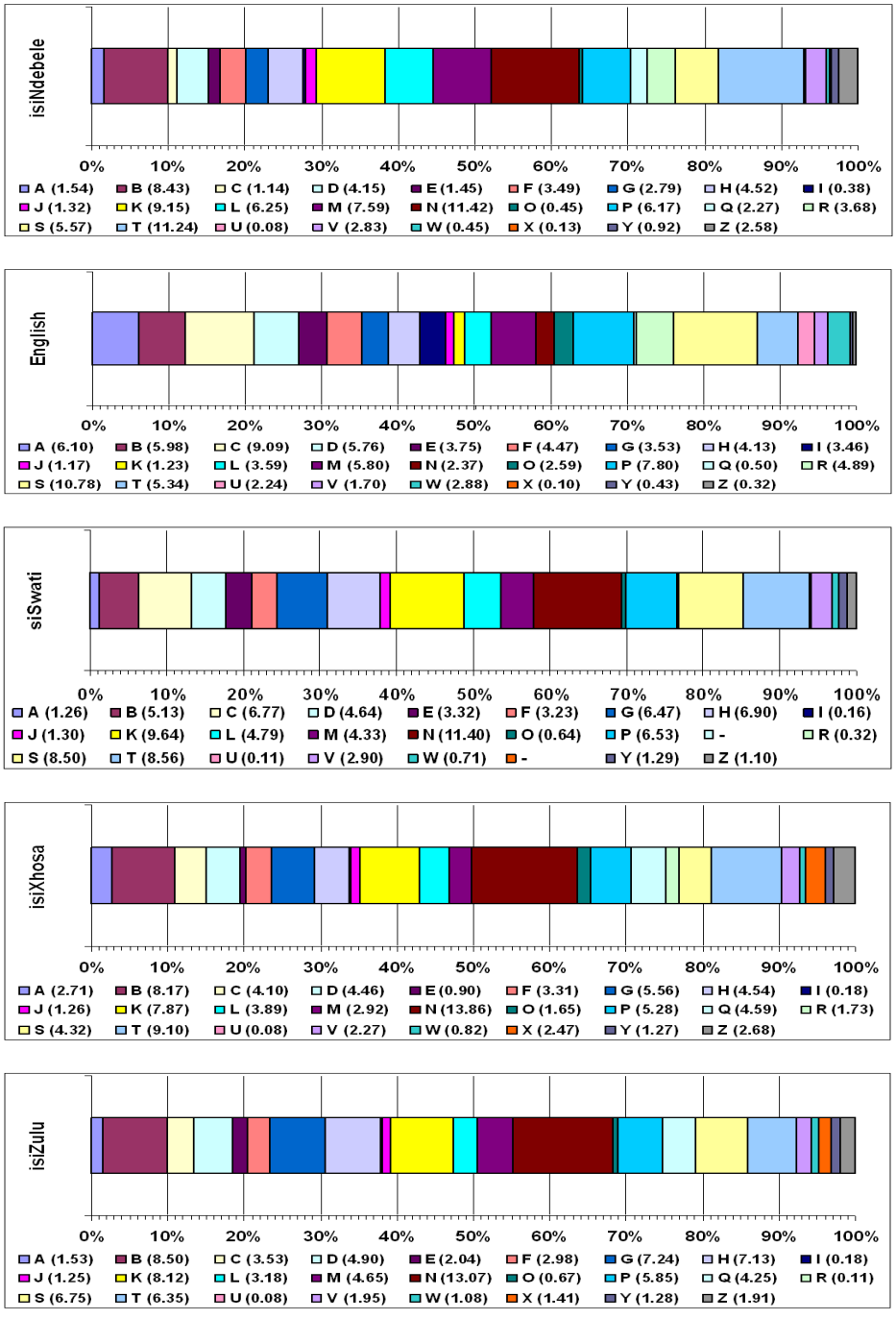
UP Strategiese Plan. 2002–2005. http://www.up.ac.za

Woordeboek van die Afrikaanse Taal. Deel XII (P–Q). 2005. Botha, W.F. (Hoofred.). Stellenbosch: Buro van die WAT.

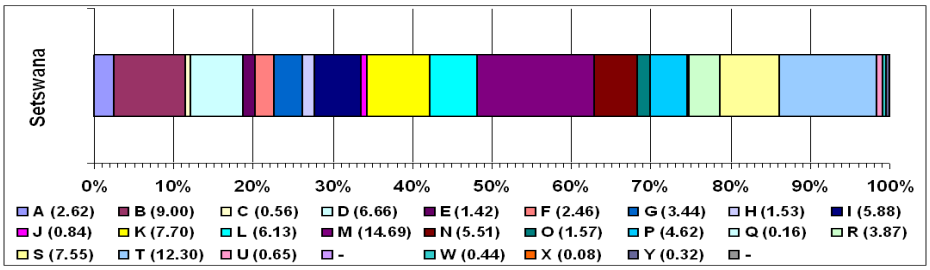
Woordeboek van die Afrikaanse Taal, Deel XIII (R). 2009. Botha, W.F. (Hoofred.). Stellenbosch: Buro van die WAT.

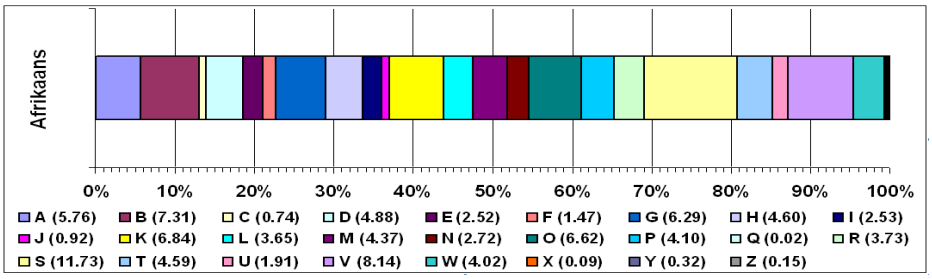
http://lexikos.journals.ac.za
doi: 10.5788/20-0-149

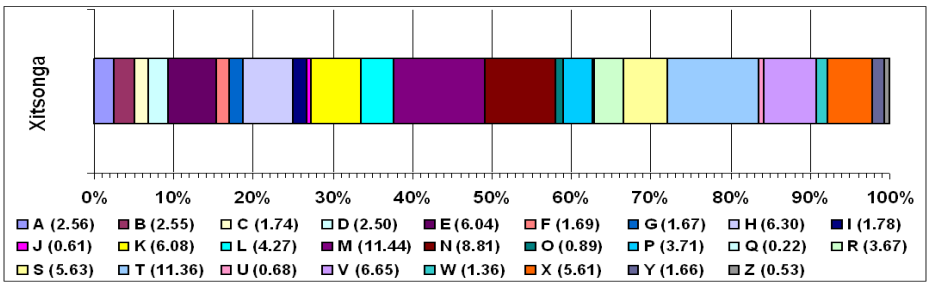
Bylaag A: Liniale vir die 11 amptelike landstale (Prinsloo en De Schryver 2005)

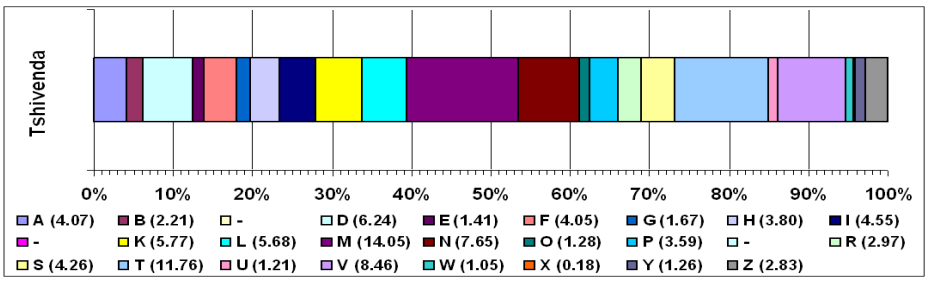


http://lexikos.journals.ac.za
doi: 10.5788/20-0-149









http://lexikos.journals.ac.za
doi: 10.5788/20-0-149