ISSN 0254–2757

QI QUÆSTIONES INFORMATICÆ

Volume 6 • Number 2 September 1988

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

QUÆSTIONES INFORMATICÆ

The official journal of the Computer Society of South Africa and of the South African Institute of Computer Scientists

Die amptelike vaktydskrif van die Rekenaarvereniging van Suid-Afrika en van die Suid-Afrikaanse Instituut van Rekenaarwetenskaplikes

Subscriptions

The annual subscription is

	SA	US	UK
Individuals	R20	$ 7	£ 5
Institutions	R30	$14	£10

to be sent to:
Computer Society of South Africa
Box 1714 Halfway House 1685

Quæstiones Informaticæ is prepared by the Computer Science Department of the University of the Witwatersrand and printed by Printed Matter, for the Computer Society of South Africa and the South African Institute of Computer Scientists.

Rekenaar Spraaksintese: Die Omskakeling van Teks na klank – 'n Prestasiemeting

M J Wagener en G de V de Kock
Departement Rekenaarwetenskap, Universiteit van Port Elizabeth, Posbus 1600, Port Elizabeth, 6000

Opsomming

Verslag word gelewer oor 'n prestasiemeting van 'n stel letter-na-klank reëls vir die omskakeling van Afrikaanse teks na foneme. Die prestasie van die stel reëls word getoets deur gebruik te maak van 'n verteenwoordigende steekproef van 1 427 Afrikaanse woorde. Verskeie statistieke wat die sukses, al dan nie, van die stel reëls weerspieël, word verskaf en bespreek.

Ontvang Julie 1988, Aanvaar Augustus 1988

1. Inleiding

Die ontwerp en implementasie van 'n stel reëls vir die omskakeling van Afrikaanse teks na foneme is bespreek in [3]. Verslag word nou gelewer oor die prestasiemeting van die stel reëls.

2. Toetsdata vir Meting

Die Nasionale Buro vir Opvoedkundige en Maatskaplike Navorsing het in Oktober 1958 'n Afrikaanse woordtelling gepubliseer [2]. Hierdie steekproef van 524 709 lopende woorde (elke voorkoms van 'n spesifieke woord word getel) bevat 22 126 unieke woorde (elke woord word eenmaal getel) en is onttrek uit 69 verskillende bronne. Die bronne is uit agt verskillende afdelings geneem en word in Tabel 1 uiteengesit.

Afdeling	Aantal bronne
Algemene tydskrifte	13
Tydskrifte vir kinders	1
Kinderlektuur	1
Boeke	18
Dagleêrs van staatsdepartemente	19
Koerante	8
Briewe, vorms en verslae	5
Bybel en kerkblaaie	4

Tabel 1

Die woorde word aangegee in volgorde van frekwensie en is alfabeties gesorteer in elke frekwensie-groep. Aangesien die steekproef so 'n wye spektrum van die Afrikaanse literatuur dek en die steekproef uitermate groot is, kan daar met sekerheid aanvaar word dat dit nie slegs 'n verteenwoordigende versameling van 'n gemiddelde woordeskat is nie, maar ook 'n verteenwoordigende versameling van die Afrikaanse woordeskat.

Die statistieke in Tabel 2 gee meer informasie oor die aard van die steekproef.

Frekwensie-interval	Unieke woorde		Lopende woorde	
	Aantal	%	Aantal	%
40317 - 71	727	3,3	401 196	76,5
70 - 61	132	0,6	8 651	1,6
60 - 51	170	0,8	9 384	1,8
50 - 41	220	1,0	9 965	1,9
40 - 31	391	1,8	13 763	2,6
30 - 21	671	3,0	16 546	3,2
20 - 11	1 380	6,2	20 152	3,8
10 - 1	18 435	83,3	45 052	8,6
Totaal	22 126	100,0	524 709	100,0

Tabel 2

Die gegewens dui aan dat die 727 mees algemene woorde in die steekproef ongeveer driekwart van die gewig van die totale steekproef dra. Verder het 83,3% van die unieke woorde 'n frekwensie minder of gelyk aan 10 en dra slegs 8,6% van die gewig. Hierdie twee uiterstes maak dit baie moeilik om 'n verdere onttrekking uit hierdie steekproef te maak. Vir die doel van die prestasiemeting is twee versamelings woorde onttrek en vervolgens sal daarna verwys word as steekproef 1 en steekproef 2 en ook later na steekproef 3, die kombinasie van die eerste twee. Steekproef 1 bestaan uit die 727 woorde van frekwensie-interval 1. Steekproef 2 bestaan uit 7 groepe van 100 woorde elk. Die groepe van 100 woorde is ewekansig getrek uit elk van die sewe oorblywende frekwensie-intervalle.

Om die transkripsie-sukses van die stel reëls te bepaal, is al die woorde van die steekproewe met hul korrekte transkripsies soos aangegee in die Uitspraakwoordeboek van Afrikaans deur Le Roux en

Pienaar [1], in 'n woordeboekstelsel gestoor. Die transkripsies gegenereer deur die stel reëls is vergelyk met die ooreenstemmende transkripsies van die woordeboekstelsel. Indien een of meer foneme in 'n woord nie ooreenstem nie, is die woord as foutief aangeteken. Let daarop dat die sukses van die stel reëls op 'n suiwer fonetiese basis gedoen is, en nie deur na die uitspraak te luister nie. Spraak word nie alleenlik deur 'n fonetiese transkripsie bepaal nie, maar ook deur die verskynsel van ko-artikulasie. Dit sal dus meer gepas wees om 'n sintesestelsel wat ko-artikulasie toepas, te beoordeel deur na die uitspraak te luister. Die ontwikkeling van so 'n stelsel word tans onderneem.

3. Resultate Verkry van Steekproewe

Die resultate verkry deur die woorde van die drie steekproewe met die stel reëls te transkribeer, word in Tabel 3 weergegee.

	Steekproewe		
	1	2	3
Woorde			
Aantal	727	700	1427
Foute	52	104	156
Sukses (%)	92,85	85,14	89,07
Sukses (%) geweeg volgens frekwensie	96,98	83,42	93,79
Foneme			
Aantal	6607	7536	14143
Foute	132	194	326
Sukses (%)	98	97,43	97,69
Sukses (%) geweeg volgens woordfrekwensie	98,52	97,37	98,25
Aantal reëls	161	161	161

Tabel 3

Die aantal reëls sluit nie reëls in om syfers, leestekens en ander spesiale karakters soos + en *, te hanteer nie.

3.1 Steekproef 1

Die 132 foutiewe foneme, soos aangegee in Tabel 3, is almal probleme met die transkripsie van klinkers. Die foute kan geklassifiseer word onder drie tipes foute.

By die eerste tipe, waarvan daar 42 is, word 'n kort-klinkerklank gegee in plaas van 'n langklinkerklank of omgekeerd.

Die tweede tipe fout is waar nasalering ontbreek of verkeerdelik voorkom – aantal foute is 16.

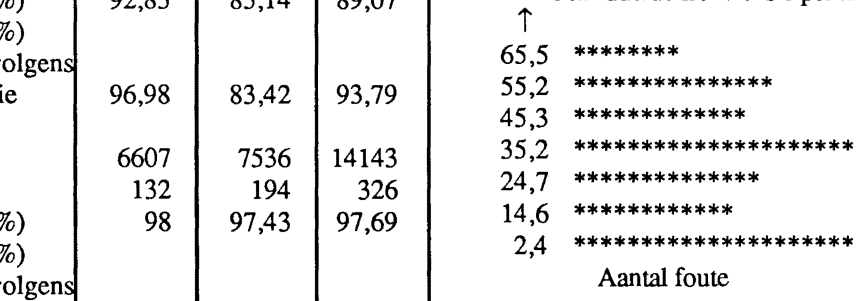
Die oorblywende 74 foute, tipe 3 foute, is ook almal klinker foute, maar is van 'n ernstiger aard, bv. "besef" word uitgespreek soos "besif".

As die kriteria vir korrekte uitspraak dus sou wees dat dit "aanvaarbaar" moet klink indien die woorde geproduseer word deur 'n spraaksintiseerder, sal die gevalle wat behoort aan die eerste twee tipe foute, nie aangeteken word nie. Die redes hiervoor is dat 'n lengte verskil in 'n klank en nasalering nie 'n groot invloed het op die verstaanbaarheid van gesintiseerde spraak nie. Dus sal 98,88 persent van die foneme vir die spesifieke steekproef aanvaarbaar klink indien dit deur 'n spraaksintiseerder geproduseer word. Van die resultate van Tabel 3 kan ook afgelei word dat daar gewoonlik meer as een foutiewe foneem in 'n foutiewe woord voorkom.

Aangesien die woordfrekwensies in hierdie steekproef nie lineêr afneem nie, is dit onmoontlik om die woorde in frekwensie-intervalle in te deel met gelyke verteenwoordiging in elke interval. Die steekproef is dus nie geskik om te bepaal of daar 'n verband tussen woordfrekwensie en transkripsie-sukses is nie.

3.2 Steekproef 2

Die resultate vir die steekproef word aangegee in Tabel 3 en die verspreiding van foute word aangedui deur Figuur 1.



Figuur 1

Die foutiewe foneme is soos voorheen geklassifiseer in drie tipes:

tipe 1 - 59
tipe 2 - 2
tipe 3 - 122
ander - 11

Daar is 11 foute wat nie aan een van die tipes behoort nie.

Soos aangedui in figuur 1 is daar 'n konstante styging in die gemiddelde frekwensie per interval en ook is daar 'n gelyke verteenwoordiging van woorde in elke interval. Die steekproef is dus geskik om vas te stel of daar 'n verband is tussen woordfrekwensie en transkripsie-sukses. Vir die doel is 'n lineêre regressielyn gepas op die data van figuur 1. Die vergelyking vir die regressielyn is

y = a + bx met
a = 19,29784 en
b = –0,12793.

Die R-waarde (regressie koëffisiënt) vir die passing is 0,5851. Hiervan kan afgelei word dat vir die data van dié steekproef, transkripsie-sukses afneem met 'n ooreenkomstige afname in woordfrekwensie. Die stel reëls is dus meer suksesvol in die hantering van 'n woordeskat waarin hoë frekwensie woorde gebruik word.

3.3 Steekproef 3

Gedurende die transkripsie van steekproef 3 is daar rekord gehou van die prestasie van elke individuele reël. Soos verduidelik in [3] bestaan daar vir elke letter of lettergreep wat getranskribeer moet word 'n groep reëls waarvan slegs een, en wel die eerste korrekte reël, gebruik word. Tellings is gehou van die aantal kere wat elke reël gebruik is.

Om 'n idee te kry van hoe 'n inperking op die aantal reëls transkripsie-sukses beïnvloed, is steekproef 3 onderwerp aan twee kleiner stelle reëls. Die eerste van hierdie stelle reëls sluit alle reëls in die oorspronklike stel wat 5 keer of meer gebruik is in en die volgende stel reëls sluit alle reëls wat 15 keer of meer gebruik is in. Die resultate word aangegee in Tabel 4.

	Alle reëls	5 of meer	15 of meer
Aantal reëls	161	109	77
Woordsukses (%) ongeweeg	89,07	86,33	77,58
Rekenaar: Tyd (sek.)	613	485	405
Geheue (grepe)	3,69K	2,99K	2,57K

Tabel 4

Die hoeveelheid geheue soos bo aangegee is slegs vir die stel reëls en nie vir enige programme nie.

Dit is duidelik dat 'n vermindering in die aantal reëls nie 'n noemenswaardige effek op die gebruik van geheue het nie, maar wel die tydsverbruik drasties verminder. Daar moet in ag geneem word dat die meeste reëls wat met die eerste vermindering weggelaat is, die reëls is wat die uitspraak van afkortings moontlik maak en dat steekproef 3 geen afkortings bevat nie. Die sukses van die kleinste stel reëls op steekproef 1 is 80,33%, en die sukses geweeg volgens woordfrekwensie is 91,3%. Gevolglik kan die stel reëls tog van groot nut wees – veral waar 'n meer alledaagse woordeskat gebruik word.

Dit is nuttig om so 'n stelsel intyds te bedryf m.a.w. die transkripsie-proses moet vinnig genoeg wees om foneme te produseer teen dieselfde tempo waarteen die spraaksintiseerder dit kan verwerk. So 'n stelsel is ge-implementeer deur gebruik te maak van die volle stel reëls. Die transkripsieproses op die betrokke mikrorekenaar is vinnig genoeg om die spraaksintiseerder te dryf en selfs 'n groter stel reëls sal nog suksesvol hanteer kan word.

4. Samevatting

Die resultate wat verkry is toon duidelik dat die meeste probleme ondervind word met die transkripsie van klinkers. Dit is veral die letter "e" wat baie foute tot gevolg het, en wat natuurlik ook baie voorkom in die Afrikaanse taal. Verskeie skommelinge en veranderings in die stel reëls het getoon dat maksimale sukses verkry is met die spesifieke algoritme. Aangesien baie foutiewe transkripsies voorkom in saamgestelde woorde, mag 'n morfologies-gebaseerde algoritme moontlik beter resultate lewer. Die koste betrokke by 'n morfologiese ontleding moet egter opgeweeg word teen die marginale verbetering wat moontlik verkry kan word. Dit verg egter verdere studie.

Bibliografie

[1] T.H. Le Roux, P. de V. Pienaar, [1976], *Uitspraakwoordeboek van Afrikaans*, J L van Schaik, Pretoria.

[2] Nasionale Buro vir Opvoedkundige en Maatskaplike Navorsing, [1958], *Afrikaanse Woordetelling*, Oktober, 1958.

[3] M.J. Wagener, [1987], Rekenaar Spraaksintese: Die Omskakeling van Teks na Klank, *Quæstiones Informaticæ*, **5** (2), 1-6.

NOTES FOR CONTRIBUTORS

The purpose of the journal will be to publish original papers in any field of computing. Papers submitted may be research articles, review articles and exploratory articles of general interest to readers of the journal. The preferred languages of the journal will be the congress languages of IFIP although papers in other languages will not be precluded.

Manuscripts should be submitted in triplicate to:

Professor J M Bishop
Department of Computer Science
University of the Witwatersrand
Johannesburg
Wits
2050

Form of manuscript

Manuscripts should be in double-space typing on one side only of sheets of A4 size with wide margins.

The first page should include the article title (which should be brief), the author's name and affiliation and address. Each paper must be accompanied by an abstract less than 200 words which will be printed at the beginning of the paper, together with an appropriate key word list and a list of relevant Computing Review categories.

Manuscripts may be provided on disc using any Apple Macintosh package or in ASCII format.

For authors wishing to provide camera-ready copy, a page specification is freely available on request from the Editor.

Tables and figures

Tables and figures should not be included in the text, although tables and figures should be referred to in the printed text. Tables should be typed on separate sheets and should be numbered consecutively and titled.

Figures should also be supplied on separate sheets, and each should be clearly identified on the back in pencil and the author's name and figure number. Original line drawings (not photocopies) should be submitted and should include all the relevant details. Photographs used as illustrations should be avoided if possible. If this cannot be avoided, glossy bromide prints are required.

Symbols

Mathematical and other symbols may be either handwritten or typewritten. Greek letters and unusual symbols should be identified in the margin. Distinction should be made between capital and lower case letters; between the letter O and zero; between the letter I, the number one and prime; between K and kappa.

References

References should be listed at the end of the manuscript in alphabetic order of the author's name, and cited in the text in square brackets. Journal references should be arranged thus:

[1] E. Ashcroft and Z. Manna, [1972], The Translation of 'GOTO' Programs to 'WHILE' programs, *Proceedings of IFIP Congress 71*, North-Holland, Amsterdam, 250-255.

[2] C. Bohm and G. Jacopini, [1966], Flow Diagrams, Turing Machines and Languages with only Two Formation Rules, *Comm. ACM*, **9**, 366-371.

[3] S. Ginsburg, [1966], *Mathematical Theory of Context-free Languages* McGraw Hill, NewYork.

Proofs

Proofs will be sent to the author to ensure that the papers have been correctly typeset and *not* for the addition of new material or major amendment to the texts. Excessive alterations may be disallowed. Corrected proofs must be returned to the production manager within three days to minimise the risk of the author's contribution having to be held over to a later issue.

Only orginal papers will be accepted, and copyright in published papers will be vested in the publisher.

Letters

A section of "Letters to the Editor" (each limited to about 500 words) will provide a forum for discussion of recent problems