

Eenentwintigste Internationale Taalkunde-Olympiade

Brasilia (Brazilië), 23–31 juli 2024

Opgave van de groepswedstrijd

Lexicostatistiek is een groep van methoden ontworpen om een inschatting te maken van hoe nauw verwant verschillende talen zijn op basis van hun vocabulaire. Deze methoden worden gewoonlijk toegepast op lange woordenlijsten door deskundigen, die handmatige annotaties toevoegen en aangeven of een bepaald woordpaar uit dezelfde bron afkomstig is. Soms passen taalkundigen echter lexicostatistische methoden toe op woordenlijsten die worden geannoteerd door geautomatiseerde procedures. Eén zo'n procedure is gebaseerd op het concept van de *medeklinkerklassen*, geïntroduceerd in 1964 door de Sovjet-Israëliëse taalkundige Aharon Dolgopolsky.

P.	p b ɓ φ β f v	K.	k g x γ q ɠ χ w	Y.	j ç (aan het begin van de wortel)	M.	m ŋ
T.	t d ɗ θ ð ʈ ɖ	R.	r r̥ ɽ ɺ l ʎ ʒ ʟ ʠ ʡ	W.	w ɯ (aan het begin van de wortel)	N.	n ɲ ɳ ɰ
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ ʑ					Q.	ʈ ɖ ʟ ʠ
H.	h ʁ ɦ ʕ ʔ h ɦ ʔ, klinkers en j ç w ɯ (behalve aan het begin van de wortel)						

Medeklinkerklassen van Dolgopolsky

Hieronder volgen geannoteerde fragmenten van woordenlijsten van meerdere taalfamilies van de wereld. De annotaties worden gegeven als cijfers in subscript. Op basis van deze lijsten zijn taalfamilie-stambomen geconstrueerd met behulp van twee vereenvoudigde versies van het zogenaamde *Starling-Nj*-algoritme, en is aan elk woord een *stabiliteitsindex* toegewezen. De bomen en stabiliteitsindices bovenaan zijn gebaseerd op handmatig geannoteerde lijsten, en die onderaan zijn gebaseerd op automatisch geannoteerde lijsten. Voor elke woordenlijst zijn er twee bomen, geconstrueerd met twee versies van het algoritme: Algoritme A en Algoritme B. Merk op dat er per woordenlijst soms meerdere bomen mogelijk zijn; in dergelijke gevallen werd één boom willekeurig geselecteerd. Aan elk knooppunt in elke boom is een lexicostatistische afstand toegewezen. Hoe groter de afstand, hoe nauwer de relatie tussen de talen. De preciezere term is dus “omgekeerde lexicostatistische afstand” eerder dan “lexicostatistische afstand”. Omwille van de eenvoudigheid wordt in deze opgave de term “lexicostatistische afstand” gebruikt.

Zowel de stabiliteitsindices als de lexicostatistische afstanden zijn afgerond op twee decimalen. Als het derde cijfer achter de komma kleiner is dan 5, rond dan naar beneden af; rond als dit niet het geval is naar boven af. Bijvoorbeeld: 2,836 wordt op 2,84 afgerond, 0,705 op 0,71 en 0,703 op 0,70. De afronding is alleen van toepassing op de waarden die aan menselijke lezers worden getoond. Met andere woorden, de computer waarop de algoritmen lopen, “ziet” de niet-afgeronde waarden.

Merk op dat het bekend is of vermoed wordt dat enkele woorden zijn ontleend aan andere talen. Zo is **jok*i*** ‘zout’ in de Kadiwéu-taal ontleend aan het Guaraní **juki**, en is **ʔa:n*i*** ‘jaar’ in het Ipai (Mesa Grande) ontleend aan het Spaanse **'ajno**.

In sommige gevallen worden meerdere synoniemen van een enkele betekenis gegeven in de woordenlijsten, gescheiden door komma's. Een voorbeeld is ‘voet’ in het Vejoz.

In de gegevens hieronder worden alle voorvoegsels gescheiden door een “=” teken en alle achtervoegsels door een “-” teken. Sommige woorden komen alleen voor met voorvoegsels. Deze woorden beginnen met een “=”-teken.

De gegevens zijn getranscribeerd in het Internationaal Fonetisch Alfabet. ' = hoofdklemtoon, ˌ = bijklemtoon (zwakker dan hoofdklemtoon), ː = lange klank, ˚ = heel korte klank, X̂Y = X en Y worden

Opgave van de groepswedstrijd

uitgesproken als een enkele klank, \acute{o} = hoge toon, \grave{o} = lage toon, \hat{o} = dalende toon, $^{\circ}o$ = voorgeglottaliseerde klank (voorafgegaan door een kortstondige blokkering van de luchtstroom in de keel), o' = ejective klank (uitgesproken door kortstondig de luchtstroom in de keel te blokkeren), o° = stemloze klank, \tilde{o} = genasaliseerde klank (uitgesproken door de neus), o^{w} = krakende stem (een diep, krassend geluid), no geeft aan dat er vóór de medeklinker wat lucht door de neus stroomt, o^h = geaspireerde medeklinker (uitgesproken met een zuchtje lucht), o^w = gelabialiseerde medeklinker (uitgesproken met afgeronde lippen), o^j = gepalatisieerde klank (uitgesproken met een deel van de tong bewogen dichtbij het harte verhemelte). **a, æ, ε, ɪ, i, ə, u, ʊ, ə, ʌ, ɒ, ɔ, ɘ, ɚ, ø** zijn klinkers. Andere speciale tekens zijn medeklinkers.

⚠ Kennis van de talen die in de opgave worden genoemd geeft geen voordeel bij het oplossen van de opgave.

Deel I. Guaicurú-taalfamilie (Argentinië, Brazilië, Paraguay)

	Toba (Oostelijk)	Pilagá	Mocoví (Chaco)	Kadiwéu
wolk	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
vuur	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
vis	njaq ₁	'nijaq ₁	naʕin ₂	nij:ogo-ḏʒegi ₃
hoofd	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
doden	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
maan	ʔawoʒojk ₁	ʔa'woʒojk ₁	ʃirajyo ₂	ep:enaj ₃
neus	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
zout	towe ₁	ol'yek ₂	ʔwe ₁	jok:i ₁
steen	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
tong	=atʃ-akət ₁	=a'tʃ-aʕət ₁	=oʔley-akən-akət ₂	=ok:el:i ₃

	Algoritme A	Algoritme B	
handmatig	<p>lexicostatistische afstand</p>		Stabiltiteitsindices: wolk 0,50 vuur 0,50 vis 0,50 hoofd 0,75 doden 1,00 maan 0,50 neus 1,00 zout 0,67 steen 0,75 tong 0,50
geautomatiseerd			Stabiltiteitsindices: wolk 0,50 vuur 0,50 vis 0,75 hoofd 0,75 doden 1,00 maan 0,50 neus 1,00 zout 0,25 steen 0,75 tong 0,50

Deel II. Nubische taalfamilie (Egypte, Soedan)

	Dongolawi	Kenuzi	Dilling	Kadaru	Debri	Birgid
doden	^h bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
maan	u ^h n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
water	^h ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
geven	^h tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
goed	^h sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
wind	^h turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
haar	^h dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
buik	^h tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
slapen	^h nɛ:r₁	ne:r₁	jer₁	dwallɛli₂	jer-i₁	ne:r-i₁
zon	^h masil₁	masil₁	ɛj₂	aju₂	ɛngal-to₃	ʔi:zi₂

	Algoritme A	Algoritme B	
handmatig			Stabiliteitsindices: doden 0,50 maan 0,83 water 1,00 geven 1,00 goed 0,50 wind 0,50 haar 0,83 buik 0,83 slapen 0,83 zon 0,50
geautomatiseerd			Stabiliteitsindices: doden 0,33 maan 0,50 water 0,50 geven 0,67 goed 0,50 wind 0,50 haar 0,83 buik 1,00 slapen 0,50 zon 0,50

- (A) (2 punten) De medeklinker **ɛ** wordt uitgesproken als de Franse *r*, aan de achterkant van de tong. Tot welke klasse behoort deze medeklinker, en hoe hebben jullie dat bepaald?
- (B) (2 punten) De Nubische boom linksboven is slechts één van de twee mogelijke bomen voor deze combinatie van het algoritme en de annotatiemethode. Teken de andere mogelijke boom.
- (C) (2 punten) De Nubische boom linksonder is slechts één van de twee mogelijke bomen voor deze combinatie van het algoritme en de annotatiemethode. Teken de andere mogelijke boom.
- (D) (2 punten) De lexicostatistische afstand 0,49, die aan de wortel van de Nubische boom rechtsboven is toegewezen, is afgerond op twee decimalen (zoals enkele andere afstanden in deze opgave). Wat is de exacte afstand?

Deel III. Mataguayische taalfamilie (Argentinië, Bolivia, Paraguay)

	Wichí (Neder- Bermejo)	Wichí (Rivada- via)	Vejoz	'Weenhayek	Iyowa'aja'	Manjui	Nivaclé (Shichaam Lhavos)	Nivaclé (Chis- hamnee Lhavos)	Maká
vuur	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	'hwat ₂	'ʔeite ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
vis	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	si'ʔjus ₋₁	ʃi'ʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
voet	=patʃ _u ₁	=qol _o ₂	=patʃ _o ₁ , =kala ₂	=pa:k'oʔ ₁	=sat ₃	=ka'laʔ ₂	=φoʔ ₄	=φoʔ ₄	=f'iʔ ₅
water	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔi'n'at ₁	ʔa'ʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
geven	=ʔweŋ _{-u} ₁	=weŋ _{-u} ₁	=ʔweŋ _{-o} ₁	=ʔweŋ _{-o} ₁	=wɛhn-aʔm ₂	=hajʔ ₃ , =wɛn ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
goed	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	'ʔes ₁	'ʔeis ₁	ʔis ₁	ʔis ₁	t=ejk'un-ej ₂
wind	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	'hlahwuʔ ₄	'hlahwu ^w ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	t'unik'i ₆
boom	haʔlo ₁	hal _o ₁	haʔla ₁	haʔlaʔ ₁	ʔa'ʔlaʔ ₁	ʔa'ʔla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
haar	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lənax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʃateʔtj ₃	=jeʔs ₄	=ʔewkux-its ₅
doden	=lon ₁	=lən ₁	=lan ₁	=la:ŋ ₁	=laʔan ₁	=lan ₁	=klan ₁	=klan ₁	=lan ₁

	Algoritme A	Algoritme B	
handmatig			Stabyliteitsindices: vuur 0,78 vis 1,00 voet 0,33 water 0,78 geven 0,44 goed 0,89 wind 0,33 boom 0,78 haar 0,67 doden 1,00
geautomatiseerd			Stabyliteitsindices: vuur 0,78 vis 0,44 voet 0,33 water 0,56 geven 0,67 goed 0,89 wind 0,22 boom 0,67 haar 0,67 doden 1,00

Deel IV. Mongoolse taalfamilie (Volksrepubliek China, Mongolië, Rusland)

(E) (10 punten) Bekijk de volgende woordenlijst. Bereken de stabiliteitsindices die met zowel de handmatige als de geautomatiseerde annotaties overeenkomen.

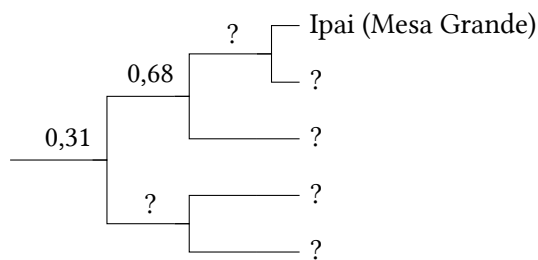
Om jullie te helpen zijn beide stabiliteitsindices voor het woord ‘alle’ al berekend. Deze indices zijn, in willekeurige volgorde, 0,36 en 0,40.

	Dagoer (Hailar)	Chamnigaans (Mantsjoe)	Boerjatisch (Khorl)	Nieuw- Bargoetisch	Ööloods	Khoshut	Kalmuks	Chalcha	Oerdoes	Oost- Joegoerisch	Bonan
alle	hɑ:₁	bölt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak₋₁	pux₃, pugt₄, xamăG₋₁	pyyyte₄, xamukʰ₋₁	tʃʰuq₅	hanə₂
schors	hails₁	qalihon₁	χoltōhōn₂	xalʰhu:₁	xolts₂	xalis₁	dursn₃	xəɣtʰōs₂	turusu₃	χalsən₁	arasun₄
buik	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwɣij₋₁	ketysy₂	ketesən₂	kele₁
vogel	dəgi₋₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendzer₂
vuur	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
weg	terg-u:l₁	qargvi₂	χargi₂, zam₋₁	zam₋₁	dzam₋₁	dzam₋₁	xɑ:-lɤə₃	tsam₋₁	tʃam₋₁	mør₄	mor₄
zout	hata:₁	dawhōn₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
zwemmen	unpa-du₁	umba₋₁	tʰamar₋₂	umb₋₁	sele₋₃	umba₋₁	us-təi₋₄, ø:m₋₅	siɣi₋₃	usu-tʃʰi-la₋₄	umpa₋₁	mba₋₁
water	ɔsə₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ʊsö₁	usun₁	qʰusun₁	sə₁
wind	kein₁	halkin₂	halxin₂	halxi₂	salxin₂	salkji₂	saljkn₂	saɣxi₂	kʰi:₁	kʰi:₁	ki₁

Deel V. Yuman-taalfamilie (Mexico, VS)

(F) (8 punten) Bekijk de volgende woordenlijst. Hieronder volgt een boom geconstrueerd op basis van dezelfde lijst. Sommige gegevens (de namen van de talen en de lexicostatistische afstanden) ontbreken. Vul de lege plekken in. Geef aan of de boom handmatig of geautomatiseerd is, en ook of deze is gegenereerd met Algoritme A of B.

	Mojave	Cocopa	Yavapai	Tipai (Jamul)	Ipai (Mesa Grande)
kort	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
vogel	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=ʃsa ₂	aʔ=ʃa ₂	ʔa:=ʃa:2
bot	n=a=s=ak ₁	'n=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
droog	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
vlees	k ^w i:k ^w ay ₁	ʔi='ma:tʃ ₂	'k ^w e:='θo-β-a ₃	'k ^w ak ₄	kuk ^w a:j-p ₁
nek	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:=puk ₂
zien	i=ju:-k ₁	'wi:2	'ʔu:1	'wi:w ₂	ə=wu:w ₂
staart	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
twee	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
jaar	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=ʔ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 punten) Enkele andere bomen werden gegenereerd voor de Yuman-taalfamilie, met de volgende lexicostatistische afstanden aan de wortel van elke boom (de lexicostatistische afstanden helemaal links van elke boom):

1. 0,20
2. 0,23
3. 0,24

Teken elk van deze bomen. Geef voor elk van de bomen aan of deze handmatig of geautomatiseerd is, en ook of deze is gegenereerd met Algoritme A of B.

(H) (3 punten) Twee afstanden gegeven in opdracht (G) zijn afgerond op twee decimalen: 0,23 is afgerond vanaf 0,225. Welke andere afstand is afgerond, en wat was de precieze waarde ervan?

(I) (4 punten) Verklaar hoe de stabiliteitsindices worden berekend.

(J) (5 punten) Verklaar hoe de lexicostatistische afstanden worden berekend.

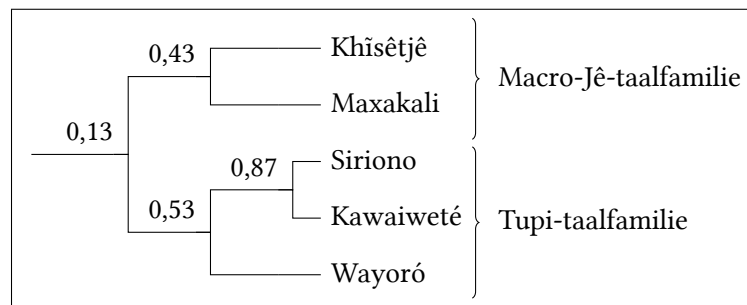
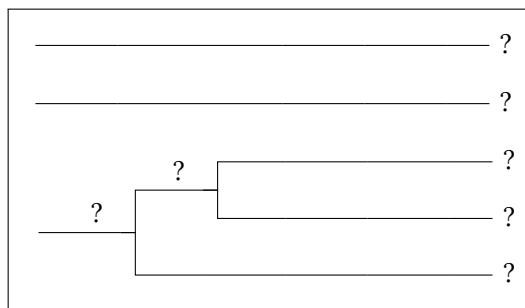
(K) (4 punten) Verklaar het verschil tussen Algoritmen A en B.

Deel VI. Macro-Jê-taalfamilie en Tupi-taalfamilie (Brazilië, Boliviaë)

(L) (28 punten) De Macro-Jê talen en Tupi talen zijn twee grote taalfamilies van Zuid-Amerika. Enkele taalkundigen veronderstellen dat ze in de verte verwant zijn. Bekijk de volgende woordenlijsten.

	A	B	Γ	Δ	E
schors	e='e-ke	h ^w ĩ='k ^h Λ	kup='pe	mĩβ̃m='tɛaj	= 'pe
buik	'e=rje	=t ^h igi	=ã'ñn	=tæj	=re'wek
bloed	e='ruki	=ka ⁿ bɾo	=d̃z=a'u	=hɛβ̃p	=ru'i
branden	'raĩ	=rɔ='k ^h ɔã	=po ^k wa	mũ=...='haβ̃p	=ra'pi
vet	e='kira	=t ^h wəmi	=d̃z=ap	=tuβ̃p	=kap
voet	'e=i	=h ^w aji	=β̃i	=pɔ'ta	=p̃i
hand	'e=o	=nĩ ^k ɔa	=β̃o	=nĩβ̃m	=pɔ
zwaar	e='usi	=wi ^t hĩ	=pɔ'ti	=β̃p'tɔj	=pɔ'ij
lever	'e=ja	= ⁿ ba	=pi'a	=tɛiβ̃pkĩ ⁿ ãj	=p̃i'ʔa
nieuw	e='jasu	= ⁿ diwi	=pa ^g op	=tiβ̃p	=pia'u
wortel	e='rao	=ja ^ɾ e	kup=kujɔ'pe	mĩβ̃m=nĩβ̃m=tɛa'tiə	=ra'pɔ
huid	'e=i	=k ^h Λ	= 'pe	=tɛaj	= 'pit
staart	e='rokoi	= ⁿ bi	=d̃z=o ^k waj	=nã='kiβ̃p	= 'raj
wit	'e=fi	=ja ^k ha	=d̃zi ^r a	=β̃p'douɥ	= 'siŋ
vleugel	e='heo	=ja ^ɾ a	=pe'o	=nĩ ^m ãuɥ	=pe'pɔ, =j̃i'wa

Hieronder volgen twee bomen geconstrueerd op basis van dezelfde lijsten. Sommige gegevens (de namen van de talen en de lexicostatistische afstanden) ontbreken. Vul de lege plekken in. Geef voor elk van de bomen aan of deze handmatig of geautomatiseerd is, en ook of deze is gegeneerd met Algoritme A of B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ In deze opdracht zijn de handmatige annotaties en stabiliteitsindices opzettelijk weggelaten.

(M) (10 punten) Geautomatiseerde procedures op basis van Dolgopolsky-klassen kunnen onjuiste resultaten opleveren. In dit voorbeeld detecteert de geautomatiseerde procedure meer overeenkomsten tussen het Siriono en een bepaalde Macro-Jê taal (Khîsêtjê) dan tussen het Siriono en andere Tupi talen. Stel een gewijzigde geautomatiseerde procedure voor die een juiste classificatie zou opleveren, als deze op de Macro-Jê en Tupi woordenlijsten boven zou worden toegepast, en beschrijf deze procedure *kort*.

⚠ Deze opdracht krijgt alleen een score als de beste teams gelijk staan.

De auteurs danken Alejandra Vidal, Maria Konoshenko, Ilya Gruntov en Jamthô Suyá voor het beantwoorden van hun vragen over specifieke talen. —*Andrey Nikulin, Milena Veneva*

Redactie: Ivan Derzhanski (technisch redacteur), Hugh Dobbs, Stanislav Gurevich, Boris Iomdin, Eimear McKnight, Andrey Nikulin (hoofdredacteur), Aleksejs Peguševs, Jan Petr, Alexander Piperski, Maria Rubinstein, Milena Veneva, Elysia Warner.

Nederlandse tekst: Elysia Warner.

Succes!