

Tjugoförsta internationella lingvistikolympiaden

Brasília (Brasilien), 23–31 juli 2024

Uppgift för lagtävlingen

Lexikostatistik är en grupp metoder som har skapats i syfte att kunna uppskatta hur närbesläktade olika språk är baserat på deras lexikon. Dessa metoder tillämpas normalt sett på långa ordlistor som har blivit manuellt annoterade av experter som indikerar vilka ordpar som tros ha ett delat ursprung. Ibland tillämpar däremot lingvister lexikostatistiska metoder på ordlistor annoterade genom automatiserade procedurer. En sådan procedur är baserad på ett koncept av s.k. *konsonantklasser*, introducerat av den sovjetisk-israeliske lingvisten Aharon Dolgopolsky år 1964.

P.	p b ɓ φ β f v	K.	k g x γ q ɠ χ ɰ	Y.	j ç (i början av en rot)	M.	m ɱ
T.	t d ɗ θ ð ʈ ɖ	R.	r ɾ ɽ l ɭ ʎ ʟ	W.	w ɰ (i början av en rot)	N.	n ɲ ɳ ɰ
S.	s z ʃ ʒ ʂ ʐ c ɟ					Q.	ʈ ɖ
H.	h ʕ ɦ ʕ ʔ h ɦ ʔ, vokaler, och j ç w ɰ (utom i början av en rot)						

Dolgopolskys konsonantklasser

Nedan finns annoterade delar av ordlistor från flera språkfamiljer i olika delar av världen. Annotationerna ges med nedsänkta siffror. Baserat på dessa listor har språkfamiljesläkträd konstruerats med två förenklade metoder av den s.k. *StarlingNj*-algoritmen, och ett *stabilitetsindex* har tilldelats varje ord. Träden och stabilitetsindexen högst upp är baserade på manuellt annoterade ordlistor, och de längst ner är baserade på listor som har blivit automatiskt annoterade. Det finns två konstruerade träd för varje ordlista, enligt två versioner av algoritmen: Algoritm A och Algoritm B. Notera att i vissa fall finns flera möjliga träd motsvarande en ordlista; i sådana fall har ett träd blivit slumpmässigt utvalt. Varje nod på varje träd har tilldelats ett lexikostatistiskt avstånd. Ju större avstånd, desto närmre besläktade är språken till varandra. Därför skulle ”inverterat lexikostatistiskt avstånd” vara en mer exakt term än ”lexikostatistiskt avstånd”. För enkelhetens skull används termen ”lexikostatistiskt avstånd” i den här uppgiften.

Både stabilitetsindexen och de lexikostatistiska avstånden är avrundade till två decimaler. Om den tredje siffran efter kommatecknet är mindre än 5, avrunda nedåt; annars, avrunda uppåt. Exempelvis avrundas 2,836 till 2,84, 0,705 till 0,71, och 0,703 till 0,70. Avrundningen gäller bara för värdena som visas för mänskliga läsare. Alltså ”ser” datorn som kör algoritmen enbart värdena utan avrundning.

Märk att vissa ord har eller tros ha blivit inlånade från andra språk. Till exempel är kadiwéu-språkets ord **jok:i** ’salt’ inlånat från guaraní **juki**, och **?a:nⁱ** ’är’ i ipai (Mesa Grande) är inlånat från spanska **’ajo**.

I vissa fall anges flera synonymer för en enskild betydelse i ordlistorna, avgränsade med komma-tecken. Ett exempel är ’fot’ på vejoz.

I datan nedan avgränsas alla prefix med tecknet ”=”, och alla suffix med tecknet ”-”. Vissa ord används bara med prefix. Dessa börjar med tecknet ”=”.

Datan är transkriberad med det internationella fonetiska alfabetet. ^ˈ = huvudbetoning, _ˌ = bibetoning (svagare än huvudbetoning), ː = långt ljud, ˚ = väldigt kort ljud, \widehat{XY} = X och Y uttalas som ett ljud, ˆ = hög ton, ˘ = låg ton, ˆ = fallande ton, ʔ = preglottaliserat ljud (föregås av en kort blockering av luftflödet i strupen), ʼ = ejektivt ljud (uttalas genom att blockera luftflödet i strupen), ˚ = tonlöst ljud, ˚ = nasaliserat ljud (uttalas genom näsan), ˚ = knarrande uttal (uttalas lågt och raspigt), ˚˚ indikerar nasalitet före konsonanten, ˚^h = aspirerad konsonant (uttalas med en luftstöt), ˚^w = labialiserad

konsonant (uttalas med rundade läppar), ɔ^j = palataliserat ljud (uttalas med en del av tungan höjd mot gommen). **a, æ, ε, ɪ, i, ɔ, ʊ, u, ə, ʌ, ɒ, ɔ, ɤ, ɟ, ɐ, ø** är vokaler. Andra specialtecken är konsonanter.

⚠ Kunskap om språken nämnda i uppgiften ger inte någon fördel i lösandet av uppgiften.

Del I. Guaicuru-språk (Argentina, Brasilien, Paraguay)

	toba (östlig)	pilagá	mocoví (Chaco)	kadiwéu
moln	l=ʔok ₁	ʔo=ʔok ₁	naweyelek ₂	lol:adi ₃
eld	nodek ₁	ʔd=oleʔ ₂	norek ₁	n=ol:edi ₂
fisk	njaq ₁	ʔnijaq ₁	naʎin ₂	nij:ogo-ɖʒegi ₃
huvud	=qajk ₁	ʔqajk ₁	=qaik ₁	=ak:ilo ₂
att döda	=alawat ₁	=aʔla:t ₁	=alawat ₁	=el:owadi ₁
måne	ʔawoʒojk ₁	ʔaʔwoʒojk ₁	ʃirajɣo ₂	ep:enaj ₃
näsa	=mik ₁	ʔmik ₁	=mik ₁	=m:iq:o ₁
salt	towe ₁	olʔyek ₂	ʔwe ₁	jok:i ₁
sten	qaʔ ₁	ʔqaʔ ₁	qaʔ ₁	wet:iga ₂
tunga	=aʔʃ-ʌʒat ₁	=aʔʃ-ʌʒat ₁	=oʔley-ʌʒan-ʌʒat ₂	=ok:eli:3

	Algorithm A	Algorithm B	
manuell	<p>lexikostatistiskt avstånd</p>		Stabilitetsindex: moln 0,50 eld 0,50 fisk 0,50 huvud 0,75 att döda 1,00 måne 0,50 näsa 1,00 salt 0,67 sten 0,75 tunga 0,50
			Stabilitetsindex: moln 0,50 eld 0,50 fisk 0,75 huvud 0,75 att döda 1,00 måne 0,50 näsa 1,00 salt 0,25 sten 0,75 tunga 0,50

Del II. Nubiska språk (Egypten, Sudan)

	dongolawi	kenuzi	dilling	kadaru	debri	birgid
att döda	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
måne	u'n-at-t₁	an-at-ti₁	nən-ti₁	nən-tu₁	nən-to₁	ma:l₂
vatten	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
att ge	'tir₁	tir₁	ti₁	ti₁	ti₁	te:n₁
god	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
vind	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
hår	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
mage	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
att sova	'nɛ:r₁	ne:r₁	ɟer₁	dwallɛli₂	ɟer-i₁	ne:r-i₁
sol	'masil₁	masil₁	ɛɟ₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	Algorithm A	Algorithm B	
manuell			Stabilitetsindex: att döda 0,50 måne 0,83 vatten 1,00 att ge 1,00 god 0,50 vind 0,50 hår 0,83 mage 0,83 att sova 0,83 sol 0,50
automatiserad			Stabilitetsindex: att döda 0,33 måne 0,50 vatten 0,50 att ge 0,67 god 0,50 vind 0,50 hår 0,83 mage 1,00 att sova 0,50 sol 0,50

- (A) (2 poäng) Konsonanten **ɟ** uttalas som ett sydsvenskt tungrots-*r*. Vilken dolgopolskygrupp tillhör den, och hur vet ni det?
- (B) (2 poäng) Det nubiska trädet längst upp till vänster är bara ett av två möjliga träd för den kombinationen av algoritm och annoteringstyp. Rita det andra möjliga trädet.
- (C) (2 poäng) Det nubiska trädet längst ned till vänster är bara ett av två möjliga träd för den kombinationen av algoritm och annoteringstyp. Rita det andra möjliga trädet.
- (D) (2 poäng) Det lexikostatistiska avståndet 0,49 (tilldelat till roten av det nubiska trädet högst upp till höger) har avrundats till två decimaler, likt vissa andra avstånd i uppgiften. Vad är det exakta avståndet?

Del III. Mataguayiska språk (Argentina, Bolivia, Paraguay)

	wichí (nedre bermejeño)	wichí (ri-vadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaklé (shichaam lhavos)	nivaklé (chisham-nee lhavos)	maká
eld	ʔitoχ ₁	ʔitəχ ₁	ʔitah ₁	ʔi:taχ ₁	ʰwat ₂	ʔeĩe ₁	ʔitaχ ₁	ʔitaχ ₁	feʔt ₂
fisk	ʔwaha ₁	waha ₁	waha ₁	ʔwa:ha ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
fot	=patʃ _{u1}	=qəɓ ₂	=patʃ _{o1} , =kala ₂	=pa:kʼoʔ ₁	=ʼsat ₃	=kaʼlaʔ ₂	=φoʔ ₄	=φoʔ ₄	=fʼiʔ ₅
vatten	ʔinot ₁	ʔinət ₁	wah ₂	ʔina:t ₁	ʔiʼnʔat ₁	ʔaʼnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
att ge	=ʔwenɔ-u ₁	=wenɔ-u ₁	=ʔwenɔ-o ₁	=ʔwenɔ-oʔ ₁	=ʼwehn-aʔm ₂	=ʼhajʔ ₃ , =ʼwen ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
god	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʼun-ej ₂
vind	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʰhlahwuʔ ₄	ʰhlahwu ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʼunikʼi ₆
träd	haʔlo ₁	halə ₁	haʔla ₁	haʔlaʔ ₁	ʔaʼlaʔ ₁	ʔaʼla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
hår	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenaχ ₂	=ʔwole ₁	=ʔwole-j ₁	=ʃateʔtʃ ₃	=jeʔs ₄	=ʔewkux-its ₅
att döda	=lon ₁	=lən ₁	=lan ₁	=la:ŋ ₁	=ʼlaʔan ₁	=ʼlan ₁	=klan ₁	=klan ₁	=lan ₁

Del IV. Mongoliska språk (Folkrepubliken Kina, Mongoliet, Ryssland)

(E) (10 poäng) Undersök följande ordlista. Beräkna stabilitetsindex motsvarande både den manuella och den automatiserade annoteringen.

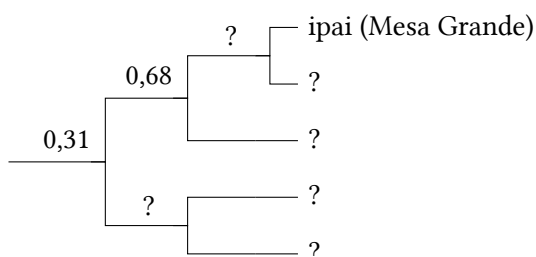
För att hjälpa er har vi redan beräknat båda stabilitetsindex för ordet 'alla'. I godtycklig ordning är dessa 0,36 och 0,40.

	dagur (hailar)	khamnigan (manchu)	burjatiska (khor)	ny- bargutiska	ööld	khoshut	kalmuck- iska	khalkha	ordos	öst- yughur	bonan
alla	hə:₁	bolt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamäg-₁	pyyyte₄, xamuk ^h -₁	tʃ ^h uq₅	hanə-₂
bark	hails₁	qalihön₁	χoltöhön₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəɣt ^h ös₂	turusu₃	χalsən₁	arasun₄
mage	ke:li₁	gətəhün₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwɣij-₁	ketysy₂	ketesən₂	kele₁
fågel	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃör₂	bendzer₂
eld	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
väg	terg-u:l₁	qargöi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-lkə₃	tsam-₁	tʃam-₁	mør₄	mor₄
salt	hata:₁	dawhön₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsä₂	taβusu₂	ta:psən₂	dabsun₂
att simma	unpa-du₁	umba-₁	t ^h amar-₂	umb-₁	sele-₃	umba-₁	us-təi-₄, ø:m-₅	siɣi-₃	usu-tʃ ^h i-la-₄	umpa-₁	mba-₁
vatten	əsə₁	oxön₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ösö₁	usun₁	q ^h usun₁	sə₁
vind	kein₁	halkin₂	halxin₂	halxi₂	salxin₂	salkʃi₂	salʃkn₂	saɣxi₂	k ^h i:₁	k ^h i:₁	ki₁

Del V. Yuma-språk (Mexiko, USA)

(F) (8 poäng) Undersök följande ordlista. Nedan visas ett träd som konstruerades baserat på samma lista. Viss data (språknamn och lexikostatistiska avstånd) saknas. Fyll i luckorna. Ange om trädet är manuellt eller automatiserat, och om det genererades med Algoritm A eller B.

	mojave	cocopa	yavapai	tiipay (Jamul)	ipai (Mesa Grande)
kort	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuŋ ₁	mə=put-k ₃
fågel	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ='ʃa ₂	ʔa:=ʃa:₂
ben	ŋ=a=s=ak ₁	'ŋ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
torr	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
kött	kʷi:kʷay ₁	'ʔi='ma:tʃ ₂	'kʷe:=ʔo-β-a ₃	'kʷak ₄	kukʷa:j-p ₁
hals	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:=ʔuk ₂	i:=puk ₂
att se	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
svans	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
två	havik-k ₁	'x=wak ₁	'hʷâk-i ₁	xə='wak ₁	xə=wak ₁
år	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃʰur-a ₃	mat-'wam ₂	ʔa:nʲ ₁



(G) (20 poäng) Några andra träd har genererats för yuman-språken, med följande lexikostatistiska avstånd vid trädens rötter (längst till vänster på varje träd):

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

Rita alla av dessa träd. För vart och ett av träden, ange om det är manuellt eller automatiserat, och om det genererades med Algoritm A eller B.

(H) (3 poäng) Två av avstånden i deluppgift (G) har avrundats till två decimaler: 0,23 har avrundats från 0,225. Vilket annat avstånd har avrundats, och vad är dess exakta värde?

(I) (4 poäng) Förklara hur stabilitetsindex beräknas.

(J) (5 poäng) Förklara hur lexikostatistiska avstånd beräknas.

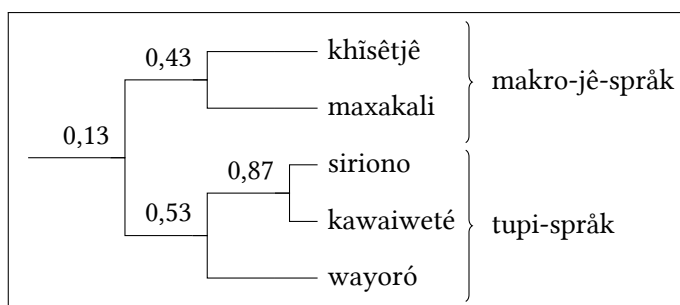
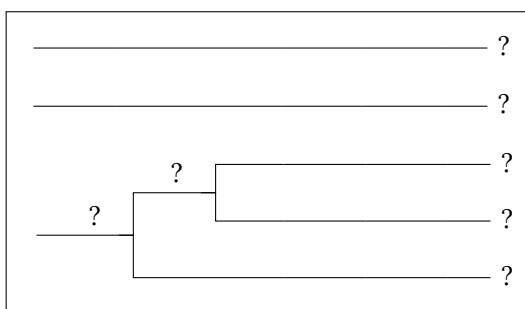
(K) (4 poäng) Förklara skillnaden mellan Algoritm A och B.

Del VI. Makro-jê-språk och tupi-språk (Brasilien, Bolivia)

(L) (28 poäng) Makro-jê och tupi är två större språkfamiljer i Sydamerika. Vissa lingvister tror att de är besläktade på stort avstånd. Undersök följande ordlistor.

	A	B	Γ	Δ	E
bark	e='e-ke	h ^w i='k ^h Λ	kup='pε	mīβm='tεaj	= 'pε
mage	'e=rje	= 't ^h igi	=ã'ün	= 'tæj	=rε'wεk
blod	e='ruki	=ka' ⁿ bɾo	=d̄z=a'u	= 'hεβp	=ru'i
att bränna	= 'raī	=rɔ='k ^h ɔ̄s	=po'k ^w a	mū=...='haβp	=ra'pɪ
fett	e='kira	= 't ^h wəmi	= 'd̄z=ap	= 'tuβp	= 'kap
fot	'e=i	= 'h ^w aji	= 'βi	=pɔ'ta	= 'pɪ
hand	'e=o	=ɲi' ^k h ^ɔ a	= 'βo	= 'ɲiβm	= 'pɔ
tung	e='usi	=wi' ^t h ^ɪ	=po'ti	=βp'təj	=pɔ'ij
lever	'e=ja	= 'nba	=pi'a	=təiβpkɪ'nāj	=pɪ'ʔa
ny	e='jasu	= 'ndiwi	=pa'gop	= 'tiβp	=pia'u
rot	e='rao	=ja'ɾe	kup=kujɔ'pε	mīβm=ɲiβm=tεa'tiə	=ra'pɔ
hud	'e=i	= 'k ^h Λ	= 'pε	= 'tεaj	= 'pit
svans	e='rokoī	= 'nbi	=d̄z=ɔ'k ^w aj	=nā:='kiβp	= 'raj
vit	'e=fī	=ja'k ^h a	=d̄zi'ra	=βp'dou	= 'sɪŋ
vinge	e='heo	=ja'ɾa	=pε'o	=ɲi'māu	=pε'pɔ, =ji'wa

Nedan visas två träd som konstruerades baserat på samma listor. Viss data (språknamn och lexi-kostatistiska avstånd) saknas. Fyll i luckorna. För vart och ett av träden, ange om det är manuellt eller automatiserat, och om det genererades med Algoritm A eller B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ De manuella annoteringarna och stabilitetsindexen har avsiktligt uteslutits från den här deluppgiften.

(M) (10 poäng) Automatiserade procedurer baserade på Dolgopolskys klasser kan leda till inkorrekta resultat. I detta exempel identifierar den automatiserade proceduren fler likheter mellan siriono och ett visst makro-jê-språk (khîsêtjê) än mellan siriono och andra tupi-språk. Föreslå en modifierad automatiserad procedur som skulle ge en korrekt klassificering när den tillämpas på makro-jê och tupi-ordlistorna, och beskriv den *kortfattat*.

⚠ Denna deluppgift kommer enbart att rättas om några av de bästa lagen annars har lika många poäng.

Författarna vill tacka Alejandro Vidal, Maria Konosjenko, Ilja Gruntov, och Jamthô Suyá för att ha besvarat frågor om specifika språk. —*Andrej Nikulin, Milena Veneva*

Redaktörer: Ivan Derzhanski (teknisk redaktör), Hugh Dobbs, Stanislav Gurevitj, Boris Iomdin, Eimear McKnight, Andrej Nikulin (chefredaktör), Aleksejs Peguševs, Jan Petr, Aleksandr Piperski, Maria Rubinstein, Milena Veneva, Elysia Warner.

Svensk text: David Hultman.

Lycka till!