

A douăzeci și una Olimpiadă internațională de lingvistică

Brasília (Brazilia), 23–31 iulie 2024

Problema pentru competiția pe echipe

Lexicostatistica este un grup de metode create pentru a estima cât de apropiate sunt diferite limbi pe baza vocabularului lor. Aceste metode se aplică în mod normal la liste lungi de cuvinte care au fost adnotate manual de către experți, care indică dacă consideră că o anumită pereche de cuvinte provine din aceeași sursă. Totuși, uneori lingviștii aplică metode lexicostatistice și la liste de cuvinte care au fost adnotate prin proceduri automate. O astfel de procedură se bazează pe conceptul de *clase consonantice*, introdus de lingvistul sovieto-israelit Aharon Dolgopolsky în 1964.

P.	p b ɓ φ β f v	K.	k g x ɣ q ɕ χ ɰ	Y.	j ɟ (la începutul rădăcinii)	M.	m ɱ
T.	t d ɗ θ ð ʈ ɖ	R.	r ɾ ɻ l ʎ ʒ ʃ ʂ	W.	w ɰ (la începutul rădăcinii)	N.	n ɲ ɳ ŋ
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ					Q.	ʈ ɖ
H.	h ʕ ɦ ʁ ʒ h ɦ ʔ, vocale și j ɟ w ɰ (exceptând la începutul rădăcinii)						

Clasele consonantice ale lui Dolgopolsky

Mai jos veți găsi adnotate fragmente ale unor liste de cuvinte din mai multe familii de limbi ale lumii. Adnotările sunt date ca cifre în indice. Pe baza acestor liste, au fost construiți arbori genealogici ai familiilor de limbi utilizând două versiuni simplificate ale așa-numitului algoritm *StarlingNJ*, iar un *indice de stabilitate* a fost atribuit fiecărui cuvânt. Arborii și indicii de stabilitate de mai sus se bazează pe liste de cuvinte adnotate manual, pe când cei de mai jos se bazează pe liste adnotate automat. Doi arbori au fost construiți pentru fiecare listă de cuvinte, utilizând două versiuni diferite ale algoritmului: Algoritmul A și Algoritmul B. N.B.: în anumite cazuri, există mai mulți arbori posibili pentru o singură listă de cuvinte; în acele cazuri, un singur arbore a fost ales în mod aleatoriu. Fiecărui nod din fiecare arbore i se atribuie o distanță lexicostatistică. Cu cât distanța este mai mare, cu atât mai apropiate sunt limbile. Un termen mai precis ar fi deci „distanță lexicostatistică inversată” în loc de „distanță lexicostatistică”. Pentru simplificare, termenul „distanță lexicostatistică” este utilizat în problemă.

Indicii de stabilitate și distanțele lexicostatistice sunt rotunjite la două zecimale. Dacă a treia zecimală este mai mică de 5, se rotunjește la valoarea inferioară; altfel, se rotunjește la valoarea superioară. Spre exemplu, 2,836 este rotunjit la 2,84, 0,705 la 0,71, iar 0,703 la 0,70. Rotunjirea se aplică doar la valorile prezentate cititorilor umani. Cu alte cuvinte, computerul care rulează algoritmul „vede” valorile nerotunjite.

N.B.: despre unele cuvinte se știe sau se bănuiește că au fost împrumutate din alte limbi. Spre exemplu, cuvântul **jok:i** ‘sare’ din limba kadiwéu este împrumutat din cuvântul guarani **juki**, iar **ʔa:nʲ** ‘an’ din ipai (de Mesa Grande) este împrumutat din cuvântul spaniol **ano**.

În unele cazuri, se dau mai multe sinonime cu același sens, separate prin virgulă. Un exemplu este ‘picior’ în limba vejoz.

În datele de mai jos, toate prefixele se separă prin semnul „=”, iar toate sufixele se separă prin semnul „-”. Unele cuvinte nu se folosesc decât cu prefixe. Acele cuvinte încep cu semnul „=”.

Datele au fost transcrise în Alfabetul Fonetic Internațional. ^ˈ = accent primar, _ˌ = accent secundar (mai slab ca accentul primar), ɔː = sunet lung, ɔ̞ = sunet foarte scurt, X̂Y = X și Y se pronunță ca un singur sunet, ˊ = ton înalt, ˋ = ton grav, ˊˋ = ton descendent, ʔˊ = sunet preglotalizat (precedat de blocarea scurtă a fluxului de aer în gât), ɔʰ = sunet ejectiv (pronunțat prin blocarea scurtă a fluxului de aer în gât), ɔ̥ = sunet surd, ɔ̃ = sunet nazalizat (pronunțat prin nas), ɔ̠ = laringalizare (sunet grav,

zgârietor), $^n\circ$ arată că aceste consoane sunt precedate de aer prin cavitatea nazală, \circ^h = consoană aspirată (pronunțată cu un suflu de aer), \circ^w = consoană labializată (pronunțată cu buzele rotunjite), \circ^j = sunet palatalizat (pronunțat cu o parte a limbii apropiindu-se de palatul dur). **a, æ, ε, i, ĩ, o, u, ʉ, ə, ʌ, ɐ, y, ø, ø** sunt vocale. Celelalte caractere speciale sunt consoane.

△ Cunoașterea limbilor menționate în problemă nu oferă niciun avantaj pentru rezolvarea problemei.

Partea I. Familia guaicuru (Argentina, Brazilia, Paraguay)

	toba (de est)	pilagá	mocoví (de Chaco)	kadiwéu
nor	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
foc	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
pește	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-ḏʒegi ₃
cap	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
a ucide	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
lună	ʔawoʃojk ₁	ʔa'woʃojk ₁	ʃirajyo ₂	ep:enaj ₃
nas	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
sare	towe ₁	ol'ʎek ₂	ʔwe ₁	jok:i ₋₁
piatră	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
limbă	=atʃ-awat ₁	=a'tʃ-aʃat ₁	=oʔley-awan-awat ₂	=ok:el:i ₃

	Algoritm A	Algoritm B	
manual	<p>distanța lexicostatistică</p>		Indici de stabilitate: nor 0,50 foc 0,50 pește 0,50 cap 0,75 a ucide 1,00 lună 0,50 nas 1,00 sare 0,67 piatră 0,75 limbă 0,50
automat			Indici de stabilitate: nor 0,50 foc 0,50 pește 0,75 cap 0,75 a ucide 1,00 lună 0,50 nas 1,00 sare 0,25 piatră 0,75 limbă 0,50

Partea a II-a. Familia nubiană (Egipt, Sudan)

	dongolă	kenuzi	dilling	kadero	debri	birked
a ucide	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
lună	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
apă	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
a da	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
bun	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
vânt	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
păr	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
burtă	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
a dormi	'nɛ:r₁	ne:r₁	ɟer₁	dwallɛli₂	ɟer-i₁	ne:r-i₁
soare	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	Algoritmul A	Algoritmul B	
manual			Indici de stabilitate: a ucide 0,50 lună 0,83 apă 1,00 a da 1,00 bun 0,50 vânt 0,50 păr 0,83 burtă 0,83 a dormi 0,83 soare 0,50
automat			Indici de stabilitate: a ucide 0,33 lună 0,50 apă 0,50 a da 0,67 bun 0,50 vânt 0,50 păr 0,83 burtă 1,00 a dormi 0,50 soare 0,50

- (A) (2 puncte) Consoana **Ɂ** se pronunță precum *r* în franceză, cu partea din spate a limbii. Cărei clase Dolgopolsky aparține și cum v-ați dat seama?
- (B) (2 puncte) Arborele nubian din stânga sus este doar unul din doi arbori posibili pentru această combinație de algoritm și tip de adnotare. Desenați celălalt arbore posibil.
- (C) (2 puncte) Arborele nubian din stânga jos este doar unul din doi arbori posibili pentru această combinație de algoritm și tip de adnotare. Desenați celălalt arbore posibil.
- (D) (2 puncte) Distanța lexicostatistică 0,49 (atribuită rădăcinii arborelui nubian din dreapta sus) a fost rotunjită la două zecimale, precum alte distanțe din problemă. Care este distanța exactă?

Partea a III-a. Familia mataco (Argentina, Bolivia, Paraguay)

	wichi (din Bermejo de Jos)	wichi (de Rivadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaklé (de Shichaam Lhavos)	nivaklé (de Chisham-nee Lhavos)	maká
foc	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	'hwat ₂	'ʔeite ₁	ʔitax ₁	ʔitax ₁	fe't ₂
pește	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	si'ʔjus ₋₁	ʃi'ʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
picior	=patʃu ₁	=qolb ₂	=patʃo ₁ , =kala ₂	=pa:k'oʔ ₁	=sat ₃	=ka'ʔlaʔ ₂	=foʔ ₄	=foʔ ₄	=f'iʔ ₅
apă	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔi'n'at ₁	ʔa'ʔnat ₁	jina't ₁	jina't ₁	iweliʔ ₃
a da	=ʔwenɔ-u ₁	=wenɔ-u ₁	=ʔwenɔ-o ₁	=ʔwenɔ-oʔ ₁	=wɛhn-aʔm ₂	=hajʔ ₃ , =wɛn ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
bun	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	'ʔes ₁	'ʔeis ₁	ʔis ₁	ʔis ₁	t=ejk'un-ej ₂
vânt	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:t ₂ , =x ^w ox ^w ₃	'hlahwuʔ ₄	'hlahwuʔ ₄	ʔaβi'm ₅	ʔaβi'm ₅	t'unik'i ₆
copac	ha'lo ₁	halo ₁	ha'la ₁	ha'laʔ ₁	ʔa'ʔlaʔ ₁	ʔa'ʔla-k ₁	ʔa'ʔkxi-juk ₂	ji'klaʔ ₁	naxka-k ₃
păr	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔate'ʔʃ ₃	=je's ₄	=ʔewkux-its ₅
a ucide	=lon ₁	=lɔn ₁	=lan ₁	=la:ŋ ₁	=laʔan ₁	=lan ₁	=kla ₁	=kla ₁	=lan ₁

	Algoritmul A	Algoritmul B	
manual			Indici de stabilitate: foc 0,78 pește 1,00 picior 0,33 apă 0,78 a da 0,44 bun 0,89 vânt 0,33 copac 0,78 păr 0,67 a ucide 1,00
automat			Indici de stabilitate: foc 0,78 pește 0,44 picior 0,33 apă 0,56 a da 0,67 bun 0,89 vânt 0,22 copac 0,67 păr 0,67 a ucide 1,00

Partea a IV-a. Familia mongolică (Republica Populară Chineză, Mongolia, Rusia)

(E) (10 puncte) Examinați următoarea listă de cuvinte. Calculați indicii de stabilitate corespunzători adnotărilor manuale și automate.

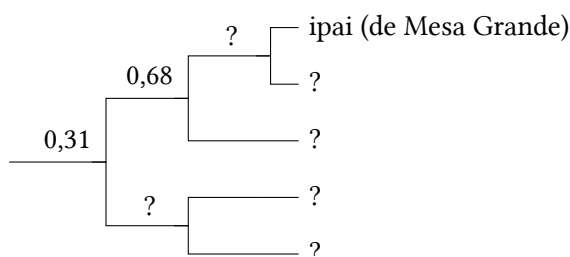
Pentru a vă ajuta, am calculat deja ambii indici de stabilitate pentru cuvântul ‘toți’. În ordine aleatorie, aceștia sunt 0,36 și 0,40.

	dagură (de Hailar)	hamnigană (manchu)	buriată (de Khori)	bargu no-uă	ööld	khoșută	kalmâcă	khalkha	ordos	yugura de est	bonan
toți	hɔ:₁	bolt₂	bɔxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pu₃, pu- gt₄, xama- G-₁	pyyyte₄, xamuk ^h -₁	tʃ ^h uq₅	hanə-₂
scoarță	hails₁	qalihon₁	χoltɔhɔn₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xɔʃt ^h ɔs₂	turusu₃	χalsən₁	arasun₄
burtă	ke:li₁	gɔtəhɔn₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwʃɟij-₁	ketysy₂	ketesən₂	kele₁
pasăre	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃɔr₂	bendzer₂
foc	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	χal₁
drum	terg-u:l₁	qargɔi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-lkə₃	tsam-₁	tʃam-₁	mør₄	mor₄
sare	hata:₁	dawhɔn₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsun₂
a înota	unpa-du₁	umba-₁	t ^h amar-₂	umb-₁	sele-₃	umba-₁	us-tɕi-₄, ø:m-₅	siɟi-₃	usu-tʃ ^h i-la-₄	umpa-₁	mba-₁
apă	ɔsɔ₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ɔsɔ₁	usun₁	q ^h usun₁	sə₁
vânt	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkʃi₂	salʃkn₂	saɟxi₂	k ^h i:₁	k ^h i:₁	ki₁

Partea a V-a. Familia yuman (Mexic, Statele Unite ale Americii)

(F) (8 puncte) Examinați următoarea listă de cuvinte. Mai jos se dă un arbore construit pe baza aceleiași liste. Unele date (numele limbilor și distanțele lexicostatistice) lipsesc. Completați spațiile libere. Precizați dacă arborele este manual sau automat, precum și dacă a fost generat utilizând algoritmul A sau B.

	mohave	cocopa	yavapai	tiipai (de Jamul)	ipai (de Mesa Grande)
scurt	wena=wen-a ₁	'xɬ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
pasăre	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=ʔʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:2
os	n=a=s=ak ₁	'n=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
uscat	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
carne	k ^w i:k ^w ay ₁	ʔi='ma:tʃ ₂	'k ^w e:=ʔo-β-a ₃	'k ^w ak ₄	kuk ^w a:j-p ₁
gât	maʎaqe ₁	'm=puk ₂	'mlq ₁	i='puk ₂	i:=puk ₂
a vedea	i=ju:-k ₁	'wi:2	'ʔu:1	'wi:w ₂	ə=wu:w ₂
coadă	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
doi	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
an	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=ʔʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 de puncte) Au fost generați câțiva alți arbori pentru familia yuman, cu următoarele distanțe lexicostatistice la rădăcina arborelui (distanțele lexicostatistice din marginea stângă a fiecărui arbore):

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

Desenați fiecare din acești arbori. Pentru fiecare arbore, specificați dacă este manual sau automat, precum și dacă a fost generat utilizând algoritmul A sau B.

(H) (3 puncte) Două din distanțele date în cerința (G) au fost rotunjite la două zecimale: 0,23 a fost rotunjit de la 0,225. Ce altă distanță a fost rotunjită și care este valoarea ei exactă?

(I) (4 puncte) Explicați cum sunt calculați indicii de stabilitate.

(J) (5 puncte) Explicați cum sunt calculate distanțele lexicostatistice.

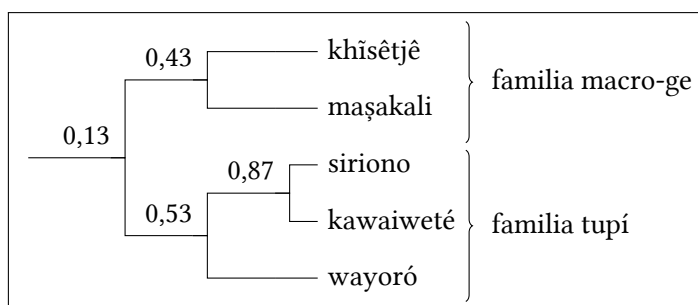
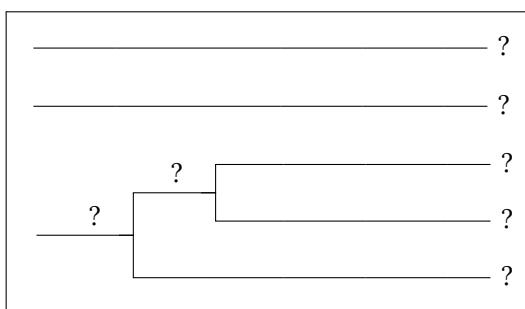
(K) (4 puncte) Explicați diferența dintre algoritmi A și B.

Partea a VI-a. Familia macro-ge și familia tupí (Brazilia, Bolivia)

(L) (28 de puncte) Macro-ge și tupí sunt două familii majore de limbi din America de Sud. Unii lingviști le consideră rude îndepărtate. Examinați următoarele liste de cuvinte.

	A	B	Γ	Δ	E
scoarță	e='e-ke	h ^w i='k ^h Λ	kup='pe	mīβm='tεaj	'pe
burtă	'e=rje	't ^h igi	=ã'ün	'tæj	=rε'wεk
sânge	e='ruki	=ka'nbrɔ	=d̥z=a'u	'hεβp	=ru'i
a arde	'rai	=rɔ='k ^h ɔ̃	=po'k ^w a	mũ=...'haβp	=ra'pi
grăsime	e='kira	't ^h wəmi	'd̥z=ap	'tuβp	'kap
picioar	'e=i	'h ^w aji	'βi	=pɔ'ta	'pi
mână	'e=o	=nī'k ^h ɔ̃	'βo	'nīβm	'pɔ
greu	e='usi	=wi't ^h i	=pɔ'ti	=βp'təj	=pɔ'ij
ficat	'e=ja	'nba	=pi'a	=tεiβpkĩ'nāj	=pi'ʔa
nou	e='jasu	'ndiwi	=pa'gop	'tiβp	=pia'u
rădăcină	e='rao	=ja'ɾe	kup=kujɔ'pe	mīβm=nīβm=tεa'tiə	=ra'pɔ
piele	'e=i	'k ^h Λ	'pe	'tεaj	'pit
coadă	e='rokoĩ	'nbi	=d̥z=o'k ^w aj	=nã:'kiβp	'raj
alb	'e=ʃi	=ja'k ^h a	=d̥zi'ra	=βp'douɥ	'sĩɲ
aripă	e='heo	=ja'ɾa	=pe'o	=nī'māuɥ	=pe'pɔ, =ji'wa

Mai jos se dau doi arbori construiți pe baza aceluiași liste. Unele date (numele limbilor și distanțele lexicostatistice) lipsesc. Completați spațiile libere. Pentru fiecare arbore, specificați dacă este manual sau automat, precum și dacă a fost generat utilizând algoritmul A sau B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ Adnotările manuale și indicii de stabilitate au fost omiși în mod intenționat în această cerință.

(M) (10 puncte) Procedurile automate bazate pe clasele Dolgopolsky pot produce rezultate incorecte. În acest exemplu, procedura automată detectează mai multe asemănări între siriono și o anumită limbă macro-ge (khîsêjtjê) decât între siriono și alte limbi tupí. Propuneți și descrieți *pe scurt* o procedură automată modificată care ar produce o clasificare corectă a limbilor macro-ge și tupí pornind de la listele de mai sus.

⚠ Această cerință va fi evaluată doar în eventualitatea unei egalități între cele mai bune echipe.

Autorii le mulțumesc Alejandrei Vidal, Mariei Konoshenko, lui Ilya Gruntov și lui Jamthô Suyá pentru răspunsurile la întrebări despre anumite limbi. —*Andrei Niculin, Milena Veneva*

Redactori: Ivan Derjanski (redactor tehnic), Hugh Dobbs, Stanislav Gurevici, Boris Iomdin, Liam McKnight, Andrei Niculin (redactor-șef), Aleksejs Peguševs, Jan Petr, Alexandr Piperski, Maria Rubinștein, Milena Veneva, Elysia Warner.

Textul în română: Dan-Mircea Mirea.

Succes!