

Двадцять перша Міжнародна олімпіада з лінгвістики

Бразилія (Бразилія), 23–31 липня 2024 року

Задача для командного змагання

Лексикостатистика — це сукупність методів, покликаних оцінити на основі лексики, наскільки близько споріднені між собою певні мови. Ці методи здебільшого застосовують до розмічених вручну експертами довгих списків слів. Для кожної пари слів експерти вказують, чи вважають вони, що ці слова походять з одного джерела. Інколи лінгвісти застосовують лексикостатистичні методи до списків слів, розмічених за допомогою автоматизованих процедур. Одна така процедура заснована на понятті *класів приголосних*, запропонованих радянсько-ізраїльським лінгвістом А. Б. Долгопольським у 1964 році.

P.	p b β f v	K.	k g x ɣ q ɕ χ ʁ	Y.	j ɟ (на початку кореня)	M.	m ɱ
T.	t d θ ð ʈ ɖ	R.	r ɾ ɻ l ʎ ʟ ʟ̥	W.	w ɱ (на початку кореня)	N.	n ɲ ɳ ɳ̥
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ̥					Q.	ʈ̥ ɖ̥
H.	ħ ʕ ɳ̥ ʔ h ɦ ʔ, голосні та j ɟ w ɱ (крім як на початку кореня)						

Класи приголосних Долгопольського

Нижче ви знайдете розмічені фрагменти списків слів для кількох мовних родин світу. Розмітку подано у вигляді нижніх індексів. Відповідно до цих списків були сконструйовані дерева мовних родин за допомогою двох спрощених варіантів так званого алгоритму *StarlingNj* і кожному слову був приписаний *індекс стабільності*. Дерева й індекси стабільності вгорі базуються на списках з ручною розміткою, а внизу — на списках з автоматизованою розміткою. Для кожного списку подано два дерева, які базуються відповідно на алгоритмах А і Б. Зверніть увагу на те, що в деяких випадках одному списку слів можуть відповідати кілька дерев; у таких випадках було обрано лише одне дерево випадковим чином. Кожному вузлу на кожному дереві приписано лексикостатистичну відстань. Чим більше її значення, тим тісніше пов'язані мови. Отже, було б точніше говорити не про “лексикостатистичну відстань”, а про “зворотну лексикостатистичну відстань”. Для спрощення в цій задачі ми використовуємо термін “лексикостатистична відстань”.

Як індекси стабільності, так і лексикостатистичні відстані округлені до двох знаків після коми. Якщо третій знак після коми менший ніж 5, значення округлюється вниз, в іншому випадку — вгору. Так, 2,836 округлюється до 2,84, 0,705 — до 0,71, а 0,703 — до 0,70. Округлення поширюється лише на значення, які показуються користувачеві. Інакше кажучи, комп'ютер у процесі застосування алгоритмів “бачить” неокруглені значення.

Зверніть увагу на те, що деякі слова — відомі чи ймовірні запозичення з інших мов. Так, слово **jok:i** ‘сіль’ мови кадівео запозичене з гуарані (**juki**), а **?a:nj** ‘рік’ мови іпай (Меса-Гранде) — з іспанської (**'año**).

Іноді для одного значення через кому подано кілька синонімів. Прикладом є ‘нога’ мовою вехоз.

У наведених нижче даних всі префікси відокремлені знаком “=”, а суфікси — знаком “-”. Деякі слова використовуються тільки з префіксами. Вони починаються зі знака “=”.

Дані записані міжнародним фонетичним алфавітом. ^ˈ = головний наголос, _ˌ = побічний наголос (слабший за головний), ː = довгий звук, ˚ = надкороткий звук, X̣Ỵ = X і Y вимовляються як один звук, ˊ = високий тон, ˋ = низький тон, ˆ = спадний тон, ˚ = преглоталізований

приголосний (якому передують коротке зімкнення гортані), \circ' = еективний приголосний (який супроводжується коротким зімкненням гортані), \circ = глухий приголосний, $\tilde{\circ}$ = носовий звук (який вимовляється в ніс), \circ = ларингалізація (низький, скрипучий звук), \circ^{h} позначає потік повітря через ніс перед приголосним, \circ^{w} = придиховий приголосний (вимовляється з придихом, різким видиханням повітря), \circ^{w} = огублений приголосний (вимовляється з округленням губ), \circ^{j} = м'який приголосний. **a, æ, ε, ɪ, i, ə, u, ɯ, ə, ʌ, ɒ, ə, ʊ, ɐ, ø** – голосні звуки. Інші особливі символи позначають приголосні.

⚠ Знання будь-якої з мов, використаних у задачі, не дає переваг для її розв'язку.

Частина I. Гвайкурська родина (Аргентина, Бразилія, Парагвай)

	тоба (східна)	пілага	мокові (Чако)	кадівео
хмара	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
вогонь	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
риба	njaq ₁	'nijaq ₁	naʃin ₂	nij:ogo-ḏʒegi ₃
голова	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
вбити	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
місяць	ʔawoʔojk ₁	ʔa'woʔojk ₁	ʃiraʒʊo ₂	eɾ:enaj ₃
ніс	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
сіль	towe ₁	ol'ʒek ₂	ʔwe ₁	jok:i ₁
камінь	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
язик	=atʃ-akat ₁	=a'tʃ-aʃat ₁	=oʔley-awan-akat ₂	=ok:el:i ₃

	алгоритм А	алгоритм Б	
ручне	<p>лекси́костатистична відстань</p>		Індекси стабільності: хмара 0,50 вогонь 0,50 риба 0,50 голова 0,75 вбити 1,00 місяць 0,50 ніс 1,00 сіль 0,67 камінь 0,75 язик 0,50
автоматизоване			Індекси стабільності: хмара 0,50 вогонь 0,50 риба 0,75 голова 0,75 вбити 1,00 місяць 0,50 ніс 1,00 сіль 0,25 камінь 0,75 язик 0,50

Частина II. Нубійська родина (Єгипет, Судан)

	донголаві	кенузі	діллінґ	кадару	дебрі	біркед
вбити	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
місяць	u'n-at-t₁	an-at-ti₁	nən-ti₁	nən-tu₁	nən-to₁	ma:l₂
вода	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
дати	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
добрий	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
вітер	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
волосся	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
живіт	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
спати	'nɛ:r₁	ne:r₁	ɟer₁	dwalleli₂	ɟer-i₁	ne:r-i₁
сонце	'masil₁	masil₁	ɛɟ₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	алгоритм А	алгоритм Б	
ручне			Індекси стабільності: вбити 0,50 місяць 0,83 вода 1,00 дати 1,00 добрий 0,50 вітер 0,50 волосся 0,83 живіт 0,83 спати 0,83 сонце 0,50
автоматизоване			Індекси стабільності: вбити 0,33 місяць 0,50 вода 0,50 дати 0,67 добрий 0,50 вітер 0,50 волосся 0,83 живіт 1,00 спати 0,50 сонце 0,50

- (A) (2 бала) Приголосний **ɟ** вимовляється як гаркаве *p*, тобто задньою частиною язика. До якого класу Долгопольського він належить і як ви це встановили?
- (B) (2 бала) Нубійське дерево у лівому верхньому кутку – лише одне з двох можливих для поданого поєднання алгоритму та типу розмітки. Намалуйте інше можливе дерево.
- (C) (2 бала) Нубійське дерево у лівому нижньому кутку – лише одне з двох можливих для поданого поєднання алгоритму та типу розмітки. Намалуйте інше можливе дерево.
- (D) (2 бала) Лексикостатистична відстань 0,49, написана біля кореня нубійського дерева у правому верхньому кутку, як і деякі інші відстані в цій задачі, округлена до двох знаків після коми. Яка точна відстань?

Частина III. Матагвайська родина (Аргентина, Болівія, Парагвай)

	вічі (нижня бермехе-ньо)	вічі (Рі-вадавія)	вехоз	ноктен	ійоґвааґа	манхуй	нівакле (шичаам хлавос)	нівакле (чишамнее хлавос)	мака
вогонь	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʰwat ₂	ʔeite ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
риба	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
нога	=patʃ _u ₁	=qol ₂	=patʃ _o ₁ , =kala ₂	=pa:kʰoʔ ₁	=ʔsat ₃	=kaʔlaʔ ₂	=foʔ ₄	=foʔ ₄	=fʔiʔ ₅
вода	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
дати	=ʔweŋ-u ₁	=weŋ-u ₁	=ʔweŋ-o ₁	=ʔweŋ-oʔ ₁	=ʔwehn-aʔm ₂	=ʔhajʔ ₃ , =ʔweŋ ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
добрий	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
вігер	ʔinwok ^w ₁	ʔinwok ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʰhlahwuʔ ₄	ʰhlahwuuʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
дерево	haʔlo ₁	hal ₁	haʔla ₁	haʔlaʔ ₁	ʔaʔlaʔ ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
волосся	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔateʔʃ ₃	=jeʔs ₄	=ʔewkux-its ₅
вбити	=lon ₁	=lon ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=klan ₁	=klan ₁	=lan ₁

	алгоритм А	алгоритм Б	
ручне			Індекси стабільності: вогонь 0,78 риба 1,00 нога 0,33 вода 0,78 дати 0,44 добрий 0,89 вітер 0,33 дерево 0,78 волосся 0,67 вбити 1,00
автоматизоване			Індекси стабільності: вогонь 0,78 риба 0,44 нога 0,33 вода 0,56 дати 0,67 добрий 0,89 вітер 0,22 дерево 0,67 волосся 0,67 вбити 1,00

Частина IV. Монгольська родина (Китайська Народна Республіка, Монголія, Росія)

(E) (10 балів) Розгляньте поданий нижче список. Розрахуйте індекси стабільності, які відповідають деревам з ручною та автоматизованою аннотаціями.

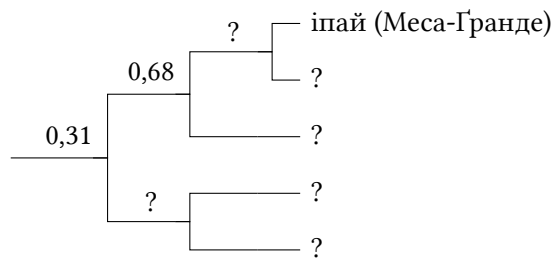
Щоб вам допомогти, ми вже розраховали обидва індекси стабільності для значення ‘всі’. Ось вони у випадковій послідовності: 0,36, 0,40.

	даурська (хайлар- ська)	хамніган- ська (мань- чжурська)	бурятська (хорин- ська)	ново- баргутська	ольотська	хошутська	калмицька	халха- монголь- ська	ордоська	східно- югурська	бонанська
всі	hə:₁	bolt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	puх₃, pugt₄, xamăġ-₁	pyyyte₄, xamukᵃ-₁	tʃᵃuq₅	hanə-₂
кора	hails₁	qalihon₁	χoltəhən₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəłtᵃᵃs₂	turusu₃	χalsən₁	arasun₄
живіт	ke:li₁	getəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitᵃs₂, xiwłjij-₁	ketysy₂	ketesən₂	kele₁
птаx	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendzer₂
вогoнь	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
дорога	terg-u:l₁	qargöi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-ləᵃ₃	tsam-₁	tʃam-₁	mər₄	mor₄
сіль	hata:₁	dawhən₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
плавати	unpa-du₁	umba-₁	tᵃamar-₂	umb-₁	sele-₃	umba-₁	us-təi-₄, ø:m-₅	siği-₃	usu-tʃᵃi-la-₄	umpa-₁	mba-₁
вода	əᵃə₁	uxən₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	usö₁	usun₁	qᵃusun₁	sə₁
вітер	kein₁	halkin₂	halxin₂	halxi₂	salxin₂	salkji₂	salkn₂	sałxı₂	kᵃi:₁	kᵃi:₁	ki₁

Частина V. Юманська родина (Мексика, США)

(F) (8 балів) Розгляньте поданий нижче список. Нижче наводимо дерево, побудоване на основі того ж списку. Деякі дані (назви мов та лексикостатистичні відстані) пропущені. Заповніть пропуски. Вкажіть, дерево збудоване за допомогою ручної чи автоматизованої розмітки, а також за допомогою алгоритму А чи Б воно згенероване.

	могавська	кокопа	явапай	типай (Хамуль)	іпай (Меса-Гранде)
короткий	wena=wen-a ₁	'xɬ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuɲ ₁	mə=put-k ₃
птах	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:₂
кістка	ɲ=a=s-ak ₁	'ɲ=j:a:k ₁	'tʃ=j:a:k-a ₁	'ak ₁	aq ₁
сухий	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
м'ясо	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
шия	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:='puk ₂
бачити	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
хвіст	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
два	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
рік	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 балів) Для юманських мов були згенеровані ще кілька дерев із зазначеними нижче лексикостатистичними відстанями в корені дерева (лексикостатистичні відстані на лівому кінці кожного дерева):

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

Намалюйте кожне з цих дерев. Для кожного дерева вкажіть, збудоване воно за допомогою ручної чи автоматизованої розмітки, а також за допомогою алгоритму А чи Б воно згенероване.

(H) (3 бала) Дві відстані, наведені в завданні (G), були округлені до двох знаків після коми: значення 0,23 отримане шляхом округлення 0,225. Яка інша відстань округлена, та яке її неокруглене значення?

(I) (4 бала) Поясніть, як обчислюються індекси стабільності.

(J) (5 балів) Поясніть, як обчислюються лексикостатистичні відстані.

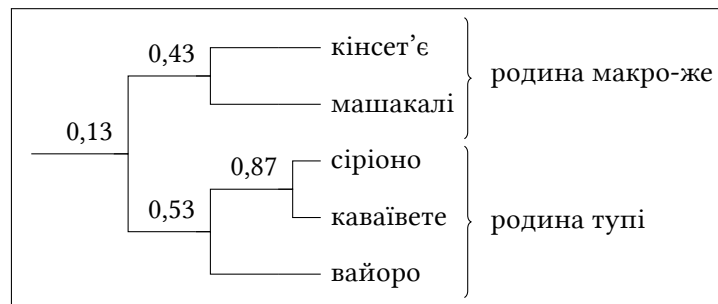
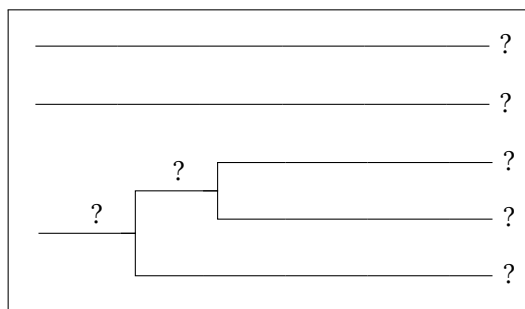
(K) (4 бала) Поясніть різницю між алгоритмами А і Б.

Частина VI. Родина макро-же і родина тупі (Бразилія, Болівія)

(L) (28 балів) Макро-же та тупі — дві великі південноамериканські родини мов. Деякі мовознавці вважають їх віддалено спорідненими. Розгляньте подані нижче списки.

	A	B	Г	Δ	E
кора	e='e-ke	h ^w i='k ^h Λ	k ^w p='pe	mīβm='tɕaj	'pe
живіт	'e=rje	't ^h igi	=ã'ñn	'tæj	=rɛ'wɛk
кров	e='ruki	=ka'n ^b ɕo	=dʒ=a'u	'hɛβp	=ru'i
палити	'raĩ	=rɔ='k ^h ɔ̃	=po'k ^w a	mũ=...'haβp	=ra'pi
жир	e='kira	't ^h wəmi	'dʒ=ap	'tuβp	'kap
нога	'e=i	'h ^w aji	'βi	=pɔ'ta	'pi
рука	'e=o	=nĩ'k ^h ɔ̃	'βo	'nĩβm	'pɔ
важкий	e='usi	=wi't ^h ĩ	=pɔ'ti	=βp'təj	=pɔ'ij
печінка	'e=ja	'nba	=pi'a	=tɕiβpkĩ'nāj	=pi'ʔa
новий	e='jasu	'ndiwi	=pa'gop	'tiβp	=pia'u
корінь	e='rao	=ja'ɕe	k ^w p=k ^w jo'pe	mīβm=nĩβm=tɕa'tiə	=ra'pɔ
шкіра	'e=i	'k ^h Λ	'pe	'tɕaj	'pit
хвіст	e='rokoi	'nbi	=dʒ=o'k ^w aj	=nã:'kiβp	'raj
білий	'e=ʃĩ	=ja'k ^h a	=dʒi'ra	=βp'douɕ	'sĩŋ
крило	e='heo	=ja'ɕa	=pe'o	=nĩ'māuɕ	=pe'pɔ, =ji'wa

Нижче наводимо два дерева, побудовані на основі тих же списків. Деякі дані (назви мов та лексикостатистичні відстані) пропущені. Заповніть пропуски. Для кожного дерева вкажіть, збудоване воно за допомогою ручної чи автоматизованої розмітки, а також за допомогою алгоритму A чи B воно згенероване.



A	B	Г	Δ	E
?	?	?	?	?

⚠ Ручна розмітка та індекси стабільності в цьому завданні навмисно не були подані.

(М) (10 балів) Автоматизовані процедури, що базуються на класах Довгопольського, можуть давати хибні результати. У цьому прикладі автоматизована процедура виявляє більше подібностей між сіріоно та певною мовою макро-же (кінсет'є), ніж між сіріоно та іншими мовами тупі. Запропонуйте свій варіант автоматизованої процедури, яка спрацювала б на наведеному матеріалі мов макро-же та тупі, і дайте її *короткий* опис.

⚠ Це завдання буде перевірятися тільки у випадку нічиєї між командами з найвищими балами.

Автори завдання дякують Алехандрі Відаль, Марії Коношенко, Іллі Грунтову та Ямто Суйя за консультації з окремих мов.
—Андрій Нікулін, Мілена Венева

Редактори: Мілена Венева, Елісія Ворнер, Станіслав Гуревич, Іван Держанський (техн. ред.),
Х'ю Доббс, Борис Іюмдін, Ліам Макнайт, Андрій Нікулін (відп. ред.), Олексій Пегушев,
Ян Петр, Олександр Піперські, Марія Рубінштейн.

Український текст: Олена Сірук.

Успіхів!