

Двадцать первая Международная олимпиада по лингвистике

Бразилиа (Бразилия), 23–31 июля 2024 г.

Задача командного соревнования

Лексикостатистика — это группа методов, призванных оценить на основе лексики, насколько близко родственны между собой те или иные языки. Как правило, эти методы применяются к длинным спискам слов, размеченным экспертами. Эксперты указывают для каждой пары, считают ли они, что эти слова восходят к одному источнику. Однако иногда лингвисты применяют лексикостатистические методы к спискам слов, размеченным с помощью автоматизированных процедур. Одна такая процедура основана на понятии *консонантных классов*, предложенных советско–израильским лингвистом А. Б. Долгопольским в 1964 году.

P.	p b β φ β f v	K.	k g x γ q c χ ц	Y.	j ç (в начале корня)	M.	m ŋ
T.	t d θ ð ð t d	R.	r r r l l t t z z l l t	W.	w м (в начале корня)	N.	n ŋ n ŋ
S.	s z ʃ z s z c z					Q.	t̪ d̪
H.	h ʏ n ʔ ʔ h h ʔ, гласные и j ç w м (кроме как в начале корня)						

Консонантные классы Долгопольского

Ниже вы найдёте размеченные фрагменты списков слов для нескольких языковых семей мира. Разметка дана в виде подписных цифр. По этим спискам были построены деревья соответствующих языковых семей с помощью двух упрощённых вариантов так называемого алгоритма *StarlingNj*, и каждому слову был приписан *индекс стабильности*. Деревья и индексы стабильности наверху основаны на списках с экспертной разметкой, а внизу — на списках с автоматизированной разметкой. При этом для каждого списка дано два дерева, основанных, соответственно, на алгоритмах А и Б. Обратите внимание на то, что в некоторых случаях есть несколько возможных деревьев, соответствующих одному списку слов; в таких случаях было отобрано только одно дерево случайным образом. Каждому узлу на каждом дереве приписано лексикостатистическое расстояние. Чем больше его значение, тем ближе связаны языки. Таким образом, было бы точнее говорить не о «лексикостатистическом расстоянии», а об «обратном лексикостатистическом расстоянии». Для простоты в этой задаче мы пользуемся термином «лексикостатистическое расстояние».

Как индексы стабильности, так и лексикостатистические расстояния округлены до двух знаков после запятой. Если третий знак после запятой менее 5, значение округляется вниз, в противном случае — вверх. Так, 2,836 округляется до 2,84, 0,705 — до 0,71, а 0,703 — до 0,70. Округление распространяется только на значения, отображаемые пользователю. Иными словами, компьютер в процессе применения алгоритмов «видит» неокруглённые значения.

Обратите внимание на то, что некоторые слова — известные или вероятные заимствования из других языков. Так, слово **jok:i** ‘соль’ языка кадивео заимствовано из гуарани **juki**, а **ʔa:nj** ‘год’ языка ипай (Меса-Гранде) — из испанского **'año**.

Иногда для одного значения через запятую перечислено несколько синонимов. В качестве примера можно назвать слова со значением ‘нога’ на языке вехос.

В приведённых ниже данных все приставки отделены знаком «=», а суффиксы — знаком «-». Некоторые слова без приставок не используются. Они начинаются со знака «=».

Данные записаны международным фонетическим алфавитом. ' = главное ударение, ˌ = побочное ударение (слабее главного), ː = долгий звук, ˚ = сверхкраткий звук, XY̆ = X и Y

произносятся, как один звук, $\acute{\circ}$ = высокий тон, $\grave{\circ}$ = низкий тон, $\hat{\circ}$ = падающий тон, $\text{?}\circ$ = прелоттализированный согласный (предварённый кратким смыканием гортани), $\text{?}'$ = эйективный согласный (сопровождается кратким смыканием гортани), $\text{?}\circ$ = глухой согласный, $\tilde{\circ}$ = назализованный звук (произносимый в нос), $\text{?}\circ$ = ларингализация (низкий, скрипучий звук), $\text{?}\circ$ обозначает поток воздуха через нос перед согласным, ?^h = придыхательный согласный (произносимый с дуновением), ?^w = огубленный согласный (произносимый с округлением губ), ?^j = мягкий согласный. **a, æ, ε, i, i, ə, u, ɥ, ə, ʌ, ɒ, ɘ, y, ø, ø** — гласные звуки. Остальные особые символы обозначают согласные.

⚠ Знание любого из использованных в задаче языков не даёт преимуществ при решении.

Часть I. Семья гуайкуру (Аргентина, Бразилия, Парагвай)

	тоба (восточный)	пилага	мокови (Чако)	кадивео
облако	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
огонь	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
рыба	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-dʒegi ₃
голова	=qajk ₁	=ʔajk ₁	=qaik ₁	=ak:ilo ₂
убить	=alawat ₁	=aʔa:t ₁	=alawat ₁	=el:owadi ₁
луна	ʔawoʎok ₁	ʔaʔwoʎok ₁	ʃirajyo ₂	ep:enaj ₃
нос	=mik ₁	=ʔmik ₁	=mik ₁	=m:iq:o ₁
соль	towe ₁	olʔyek ₂	ʔwe ₁	jok:i ₁
камень	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
язык	=aʔʃ-awat ₁	=aʔʃ-aʔat ₁	=oʔley-awan-awat ₂	=ok:el:i ₃

	алгоритм А	алгоритм Б	
экспертное	<p>↑ лексикостатистическое расстояние</p>		Индексы стабильности: облако 0,50 огонь 0,50 рыба 0,50 голова 0,75 убить 1,00 луна 0,50 нос 1,00 соль 0,67 камень 0,75 язык 0,50
автоматизированное			

Часть II. Нубийская семья (Египет, Судан)

	донголауи	кенузи	диллинг	кадару	дебри	биргид
убить	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
луна	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
вода	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
дать	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
хороший	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
ветер	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
волосы	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
живот	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
спать	'nɛ:r₁	ne:r₁	jer₁	dwalleli₂	jer-i₁	ne:r-i₁
солнце	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	алгоритм А	алгоритм Б	
экспертное			Индексы стабильности: убить 0,50 луна 0,83 вода 1,00 дать 1,00 хороший 0,50 ветер 0,50 волосы 0,83 живот 0,83 спать 0,83 солнце 0,50
автоматизированное			Индексы стабильности: убить 0,33 луна 0,50 вода 0,50 дать 0,67 хороший 0,50 ветер 0,50 волосы 0,83 живот 1,00 спать 0,50 солнце 0,50

- (А) (2 балла) Согласный **ɸ** произносится как картавое *p*, то есть задней частью языка. К какому классу Долгопольского он относится и как вы это установили?
- (В) (2 балла) Нубийское дерево в левом верхнем углу — только одно из двух возможных при заданном сочетании алгоритма и типа разметки. Нарисуйте другое возможное дерево.
- (С) (2 балла) Нубийское дерево в левом нижнем углу — только одно из двух возможных при заданном сочетании алгоритма и типа разметки. Нарисуйте другое возможное дерево.
- (D) (2 балла) Лексикостатистическое расстояние 0,49 при корне нубийского дерева в правом верхнем углу, как и некоторые другие расстояния в этой задаче, округлено до двух знаков после запятой. Каково точное расстояние?

Часть III. Матагвайская семья (Аргентина, Боливия, Парагвай)

	уичи (нижний бермехеньо)	уичи (Ривадавия)	вехос	уэхнаек	ийохуааха	манхуй	нивакле (шичаам хлавос)	нивакле (чишам-нээ хлавос)	мака
огонь	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:taχ ₁	ʔh ^h wat ₂	ʔe ^h it'e ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
рыба	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
нога	=patʃu ₁	=qolɔ ₂	=patʃo ₁ , =kala ₂	=pa:kʔoʔ ₁	=ʔsat ₃	=kaʔlaʔ ₂	=foʔ ₄	=foʔ ₄	=fʔiʔ ₅
вода	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnʔat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
дать	=ʔweŋ-u ₁	=weŋ-u ₁	=ʔweŋ-o ₁	=ʔweŋ-oʔ ₁	=ʔweh ^h n-aʔm ₂	=ʔhajʔ ₃ , =ʔweŋ ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
хороший	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
ветер	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔh ^h lahwuʔ ₄	ʔh ^h lahwuʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
дерево	haʔlo ₁	halɔ ₁	haʔla ₁	haʔlaʔ ₁	ʔaʔlaʔ ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔk ^h laʔ ₁	naxka-k ₃
волосы	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenaχ ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔateʔtʃ ₃	=jeʔs ₄	=ʔewkux-its ₅
убить	=lon ₁	=lɔn ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=k ^h lan ₁	=k ^h lan ₁	=lan ₁

	алгоритм А	алгоритм Б	
экспертное			Индексы стабильности: огонь 0,78 рыба 1,00 нога 0,33 вода 0,78 дать 0,44 хороший 0,89 ветер 0,33 дерево 0,78 волосы 0,67 убить 1,00
автоматизированное			Индексы стабильности: огонь 0,78 рыба 0,44 нога 0,33 вода 0,56 дать 0,67 хороший 0,89 ветер 0,22 дерево 0,67 волосы 0,67 убить 1,00

Часть IV. Монгольская семья (КНР, Монголия, Россия)

(E) (10 баллов) Рассмотрите следующий список. Рассчитайте индексы стабильности, соответствующие деревьям с экспертной и автоматизированной разметкой.

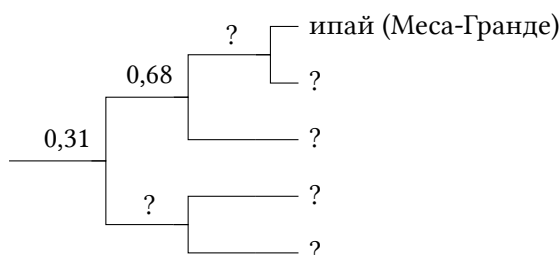
Чтобы вам помочь, мы уже рассчитали оба индекса стабильности для значения 'все'. Вот они в случайном порядке: 0,36, 0,40.

	даурский (хайларский)	хамниганский (маньчжурский)	бурятский (хоринский)	ново-баргутский	олётский	хошутский	калмыцкий	халха-монгольский	ордосский	шираюгурский	баоаньский
все	hə:₁	bolt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamăġ-₁	pyyute₄, xamukᵃ-₁	tʃᵃuq₅	hanə-₂
кора	hails₁	qalihon₁	χoltəhən₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəɮtᵃᵃs₂	turusu₃	χalsən₁	arasun₄
живот	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwɮij-₁	ketysy₂	ketesən₂	kele₁
птица	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendžer₂
огонь	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:-lkə₃	tsam-₁	tʃam-₁	mør₄	mor₄
соль	hata:₁	dawhon₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
плавать	unpa-du₁	umba-₁	tᵃamar-₂	umb-₁	sele-₃	umba-₁	us-tci-₄, ø:m-₅	siɮi-₃	usu-tʃᵃi-la-₄	umpa-₁	mba-₁
вода	əᵃə₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	öᵃᵃ₁	usun₁	qᵃusun₁	sə₁
ветер	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkji₂	salʃkn₂	saɮxi₂	kᵃi:₁	kᵃi:₁	ki₁

Часть V. Юманская семья (Мексика, США)

(F) (8 баллов) Рассмотрите следующий список. Ниже приведено дерево, построенное на основании того же списка. Некоторые данные (названия языков и лексикостатистические расстояния) пропущены. Заполните пропуски. Укажите, основано ли дерево на экспертной или автоматизированной разметке, а также построено ли оно с помощью алгоритма А или Б.

	мохаве	кокопа	явапай	типай (Хамуль)	ипай (Меса-Гранде)
короткий	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔup ₁	mə=put-k ₃
птица	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ='ʃa ₂	ʔa:=ʃa:₂
кость	ɲ=a=s=ak ₁	'ɲ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
сухой	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
мясо	k ^{wi} :k ^{way} ₁	ʔi='ma:tʃ ₂	'k ^w e:='θo-β-a ₃	'k ^w ak ₄	kuk ^w a:j-p ₁
шея	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:=puk ₂
видеть	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
хвост	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
год	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃ ⁿ ur-a ₃	mat-'wam ₂	ʔa:n ^j ₁



(G) (20 баллов) Было порождено ещё несколько деревьев для юманских языков со следующими лексикостатистическими расстояниями при корне (то есть при крайнем левом конце дерева):

- 0,20
- 0,23
- 0,24

Нарисуйте каждое из этих деревьев. Для каждого дерева укажите, основано ли оно на экспертной или автоматизированной разметке, а также построено ли оно с помощью алгоритма А или Б.

(H) (3 балла) Два расстояния, перечисленных в задании (G), были округлены до двух знаков после запятой: значение 0,23 получено путём округления 0,225. Какое другое расстояние округлено и каково его неокруглённое значение?

(I) (4 балла) Объясните, как вычисляются индексы стабильности.

(J) (5 баллов) Объясните, как вычисляются лексикостатистические расстояния.

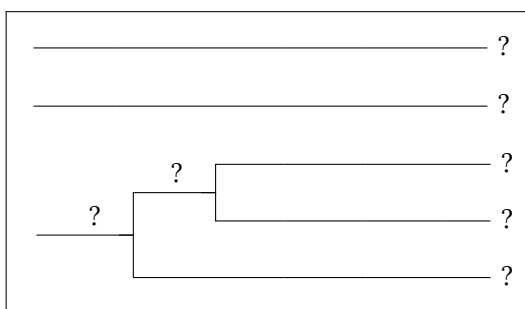
(K) (4 балла) Объясните разницу между алгоритмами А и Б.

Часть VI. Семья макро-же и тупийская семья (Бразилия, Боливия)

(L) (28 баллов) Семья макро-же и тупийская семья — крупные южноамериканские языковые семьи. Некоторые языковеды считают, что они состоят в дальнем родстве. Рассмотрите следующие списки.

	A	B	Г	Δ	Е
кора	e='e-ke	h ^w i='k ^h λ	kur='pe	mīβm='tɕaj	'pe
живот	'e=rje	't ^h igi	=ā'ūn	'təj	=re'wek
кровь	e='ruki	=ka ⁿ bɾo	=d̥z=a'u	'heβp	=ru'i
жечь	'raī	=rɔ='k ^h λɔ̃	=po'k ^w a	mū=...='haβp	=ra'pi
жир	e='kiga	't ^h wəmi	=d̥z=ap	'tuβp	'kap
нога	'e=i	'h ^w aji	'βi	=po'ta	'pi
рука	'e=o	=nī'k ^h λa	'βo	'nīβm	'pɔ
тяжёлый	e='usi	=wi't ^h i	=po'ti	=βp'təj	=pɔ'ij
печень	'e=ja	' ⁿ ba	=pi'a	=tɕiβpkī'nāj	=pi'ʔa
новый	e='jasu	' ⁿ diwi	=pa'gor	'tiβp	=pia'u
корень	e='rao	=ja'ɾe	kur=kujɔ'pe	mīβm=nīβm=tɕa'tiə	=ra'pɔ
кожа	'e=i	'k ^h λ	'pe	'tɕaj	'pit
хвост	e='rokoī	' ⁿ bi	=d̥z=ɔ'k ^w aj	=nā:'kiβp	'raj
белый	'e=fī	=ja'k ^h a	=d̥zi'ra	=βp'douɕ	'sīŋ
крыло	e='heo	=ja'ɾa	=pe'o	=nī'māuɕ	=pe'pɔ, =ji'wa

Ниже приведены два дерева, построенные на основании тех же списков. Некоторые данные (названия языков и лексикостатистические расстояния) пропущены. Заполните пропуски. Для каждого дерева укажите, основано ли оно на экспертной или автоматизированной разметке, а также построено ли оно с помощью алгоритма А или Б.



A	B	Г	Δ	Е
?	?	?	?	?

⚠ Экспертная разметка и индексы стабильности в этом задании специально были опущены.

(М) (10 баллов) Автоматизированные процедуры, основанные на классах Долгопольского, могут давать неверные результаты. В рассматриваемом примере автоматизированная процедура обнаруживает больше сходств между сирионо и одним из языков макро-же (кхинсетже), чем между сирионо и прочими тупийскими языками. Предложите свой вариант автоматизированной процедуры, которая сработала бы на приведённом материале языков макро-же и тупийских языков, и дайте её *краткое* описание.

⚠ Это задание будет проверено только в случае ничьей между командами с наивысшими баллами.

Авторы задачи благодарят Алехандру Видаль, Марию Коношенко, Илью Грунтова и Ямто Суйя за консультации по отдельным языкам. —*Андрей Никулин, Милена Венева*

Редакторы: Милена Венева, Станислав Гуревич, Иван Держанский (техн. ред.), Хью Доббс, Борис Йомдин, Лиам Макнайт, Андрей Никулин (отв. ред.), Алексей Пегушев, Ян Петр, Александр Пиперски, Мария Рубинштейн, Элисия Уорнер.

Русский текст: Андрей Никулин.

Желаем успеха!