

Vigésima primeira Olimpíada Internacional de Linguística

Brasília – DF (Brasil), 23–31 de julho de 2024

Problema da prova por equipes

A lexicoestatística é um grupo de métodos criados para avaliar quão distante é a relação entre um conjunto de línguas com base em seu vocabulário. Normalmente tais métodos são aplicados a longas listas de palavras anotadas manualmente por especialistas, que indicam se, de acordo com seu juízo, um determinado par de palavras provém de uma mesma fonte. Porém às vezes linguistas aplicam métodos lexicoestatísticos a listas de palavras anotadas por meio de procedimentos automatizados. Um desses procedimentos baseia-se no conceito de *classes consonantais*, introduzido pelo linguista soviético e israelense Aharon Dolgopolsky em 1964.

P.	p b β φ β f v	K.	k g x γ q ɣ χ ɰ	Y.	j ç (no início da raiz)	M.	m ɱ
T.	t d d̥ θ ð t̥ d̥	R.	r r̥ ʀ ɹ l ʎ ʟ ʟ̥	W.	w ɰ (no início da raiz)	N.	n ɲ ɳ ɳ̥
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ̥					Q.	ṯ ḏ
H.	ħ ʕ H ʕ̥ ʔ h ɦ ʔ, vogais e j ç w ɰ (exceto no início da raiz)						

Classes consonantais de Dolgopolsky

Abaixo vocês encontrarão fragmentos de listas de palavras de várias famílias linguísticas do mundo. As anotações são representadas por dígitos subscritos. Com base nessas listas, foram construídas árvores das respectivas famílias linguísticas por meio de duas versões simplificadas do chamado algoritmo *StarlingNj*, e foi atribuído um *índice de estabilidade* a cada palavra. As árvores e os índices de estabilidade na parte de cima baseiam-se em listas anotadas manualmente, enquanto as árvores e os índices de estabilidade na parte de baixo se baseiam em listas anotadas automaticamente. Em cada caso, há duas árvores para cada lista, construídas com versões diferentes do algoritmo: o algoritmo A e o algoritmo B. Notem que em alguns casos há múltiplas árvores possíveis que correspondem a uma lista de palavras; em tais casos, uma árvore foi escolhida aleatoriamente. Uma distância lexicoestatística é atribuída a cada nó em cada árvore. Quanto maior a distância, mais próxima a relação entre as línguas. Deste modo, o termo mais exato seria “distância lexicoestatística invertida” e não “distância lexicoestatística”. Por uma questão de simplicidade, neste problema fazemos uso do termo “distância lexicoestatística”.

Tanto os índices de estabilidade como as distâncias lexicoestatísticas foram arredondadas para duas casas decimais. Caso a terceira casa seja menor que 5, o valor é arredondado para baixo; caso contrário, é arredondado para cima. Por exemplo, 2,836 arredonda-se para 2,84, 0,705 para 0,71, 0,703 para 0,70. O arredondamento é aplicado unicamente aos valores mostrados a leitores humanos. Em outras palavras, o computador que roda os algoritmos “vê” os valores não arredondados.

Notem que algumas palavras são empréstimos conhecidos ou prováveis de outras línguas. Por exemplo, a palavra **jok:i** ‘sal’ da língua kadiwéu é um empréstimo do guarani (**juki**), ao passo que a palavra **ʔa:n̩** ‘ano’ do ‘iipay (Mesa Grande) é um empréstimo do espanhol (**‘ano**).

Às vezes são listados múltiplos sinônimos que correspondem a um único conceito, separados por uma vírgula. Um exemplo é o conceito ‘pé’ na língua vejoz.

Nos dados abaixo, todos os prefixos são separados pelo símbolo “=”, ao passo que todos os prefixos são separados pelo símbolo “-”. Algumas palavras jamais ocorrem sem prefixos. Estas vêm precedidas pelo símbolo “=”.

A transcrição dos dados segue o Alfabeto Fonético Internacional. ^ˈ = acento primário, _ˌ = acento secundário (mais fraco que o primário), ː = som longo, ˚ = som muito curto, X̄Ȳ = X e Y são pronunciados

como um único som, \acute{o} = tom alto, \grave{o} = tom baixo, \hat{o} = tom descendente, ?o = consoante preglotalizada (precedida por um curto bloqueio do fluxo de ar na glote), o' = consoante ejetiva (pronunciada com um curto bloqueio do fluxo de ar na glote), o^{v} = som desvozeado, o^{n} = som nasalizado (pronunciado pelo nariz), o^{l} = laringalização (som rouco, crepitante), o^{a} assinala um fluxo de ar através do nariz que precede ao som da consoante, o^{h} = consoante aspirada (pronunciada com um sopro), o^{w} = consoante labializada (pronunciada com os lábios arredondados), o^{j} = consoante palatalizada (pronunciada com uma parte da língua aproximando-se ao palato duro). $\text{a, \ae, \varepsilon, i, \acute{i}, \text{ɔ, u, \u00fa, \u00e3, \u00c5, \text{v, \u0254, y, \u00e9, \u00f8}}$ são vogais. Outros caracteres especiais denotam consoantes.

⚠ O conhecimento de qualquer uma das línguas mencionadas neste problema não dá vantagem para sua solução.

Parte I. Família guaicuru (Argentina, Brasil, Paraguai)

	toba (oriental)	pilagá	mocovi (chaquenho)	kadiwéu
nuvem	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
fogo	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
peixe	njaq ₁	'nijaq ₁	naʕin ₂	nij:ogo-ḏzegi ₃
cabeça	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
matar	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
lua	ʔawoʕojk ₁	ʔa'woʕojk ₁	ʕirajyo ₂	ep:enaj ₃
nariz	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
sal	towe ₁	ol'yek ₂	ʔwe ₁	jok:i. ₁
pedra	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
língua	=atʃ-aʕat ₁	=a'tʃ-aʕat ₁	=oʔley-aʕan-aʕat ₂	=ok:el:i ₃

	algoritmo A	algoritmo B	
manual	<p>distância lexicoestatística</p>		Índices de estabilidade: nuvem 0,50 fogo 0,50 peixe 0,50 cabeça 0,75 matar 1,00 lua 0,50 nariz 1,00 sal 0,67 pedra 0,75 língua 0,50
automatizada			Índices de estabilidade: nuvem 0,50 fogo 0,50 peixe 0,75 cabeça 0,75 matar 1,00 lua 0,50 nariz 1,00 sal 0,25 pedra 0,75 língua 0,50

Parte II. Família núbia (Egito, Sudão)

	dongolau	kenuzi	dilling	kadaru	debri	birgid
matar	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
lua	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
água	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
dar	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
bom	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
vento	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
cabelo	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
barriga	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
dormir	'nɛ:r₁	ne:r₁	jer₁	dwallɛli₂	jer-i₁	ne:r-i₁
sol	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	algoritmo A	algoritmo B	
manual			Índices de estabilidade: matar 0,50 lua 0,83 água 1,00 dar 1,00 bom 0,50 vento 0,50 cabelo 0,83 barriga 0,83 dormir 0,83 sol 0,50
automatizada			Índices de estabilidade: matar 0,33 lua 0,50 água 0,50 dar 0,67 bom 0,50 vento 0,50 cabelo 0,83 barriga 1,00 dormir 0,50 sol 0,50

- (A) (2 pontos) A consoante **ɣ** é pronunciada como o *r* do francês ou de algumas variedades do português de Portugal, usando a parte de trás da língua. A qual classe de Dolgopolsky ela pertence e como vocês descobriram isso?
- (B) (2 pontos) A árvore núbia no canto superior esquerdo é apenas uma de duas árvores possíveis para esta combinação do algoritmo e do tipo de anotações. Desenhem a outra árvore possível.
- (C) (2 pontos) A árvore núbia no canto inferior esquerdo é apenas uma de duas árvores possíveis para esta combinação do algoritmo e do tipo de anotações. Desenhem a outra árvore possível.
- (D) (2 pontos) A distância lexicoestatística 0,49, atribuída à raiz da árvore núbia no canto superior direito, foi arredondada para duas casas decimais, assim como outras distâncias neste problema. Qual é a distância exata?

Parte III. Família mataguaiá (Argentina, Bolívia, Paraguai)

	wichi (baixo rio Bermejo)	wichi (Rivadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaçle (shichaam lhavos)	nivaçle (chisham-nee lhavos)	maká
fogo	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʔhwat ₂	ʔeite ₁	ʔitax ₁	ʔitax ₁	fe't ₂
peixe	ʔwihat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	si'ʔjus ₋₁	ʃi'ʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
pé	=patʃ _{u1}	=qol ₂	=patʃ _{o1} , =kala ₂	=pa:k'ol ₁	=sat ₃	=ka'la ₂	=fo ₄	=fo ₄	=f'i ₅
água	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔi'njat ₁	ʔa'nat ₁	jina't ₁	jina't ₁	iweli ₃
dar	=ʔwen _{u1}	=wen _{u1}	=ʔwen _{o1}	=ʔwen _{o1}	=wehn-a _{m2}	=haj ₃ , =wen ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
bom	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejk'un-ej ₂
vento	ʔinwok ^w ₁	ʔinwok ₁	ʔihwok ^w ₁	=ja:t ₂ , =x ^w ox ^w ₃	ʔhlahwu ₄	ʔhlahwu ₄	ʔaβi'm ₅	ʔaβi'm ₅	t'unik'i ₆
árvore	ha'lo ₁	hal ₁	ha'la ₁	ha'la ₁	ʔa'la ₁	ʔa'la-k ₁	ʔa'kxi-juk ₂	ji'kla ₁	naxka-k ₃
cabelo	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenaç ₂	=ʔwole ₁	=ʔwole-j ₁	=sate'ʔ ₃	=je's ₄	=ʔewkux-its ₅
matar	=lon ₁	=lon ₁	=lan ₁	=la:ŋ ₁	=laʔan ₁	=lan ₁	=klan ₁	=klan ₁	=lan ₁

	algoritmo A	algoritmo B	
manual			Índices de estabilidade: fogo 0,78 peixe 1,00 pé 0,33 água 0,78 dar 0,44 bom 0,89 vento 0,33 árvore 0,78 cabelo 0,67 matar 1,00
automatizada			Índices de estabilidade: fogo 0,78 peixe 0,44 pé 0,33 água 0,56 dar 0,67 bom 0,89 vento 0,22 árvore 0,67 cabelo 0,67 matar 1,00

Parte IV. Família mongólica (República Popular da China, Mongólia, Rússia)

(E) (10 pontos) Examinem a seguinte lista. Calculem os índices de estabilidade que correspondem às anotações manuais e automáticas.

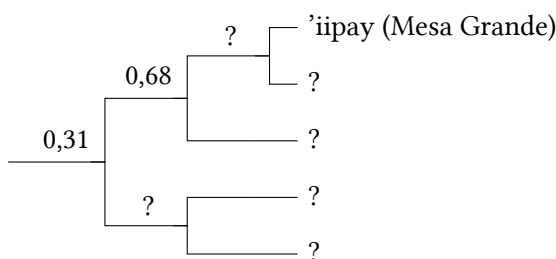
Para facilitar a tarefa de vocês, nós já calculamos os dois índices para o conceito ‘todos’. Em ordem aleatória, são esses: 0,36 e 0,40.

	daur (hai-lar)	khamnigan (manchu)	buriato (khorí)	bargut novo	ööld	khochut	calmuco	calca	ordos	chira yugur	bonã
todos	hɔː ₁	bolt ₂	buxiː ₃	bygd ₄	tsug ₅	lug ₅	tsuk ₅ , xamak ₋₁	pux ₃ , pugt ₄ , xamăġ ₋₁	pyyyte ₄ , xamuk ^h ₋₁	tʃ ^h uq ₅	hanə ₋₂
casca	hails ₁	qalihon ₁	χoltəhən ₂	xalʃhuː ₁	xolts ₂	xalis ₁	dursn ₃	xəɮt ^h ə̆s ₂	turusu ₃	χalsən ₁	arasun ₄
barriga	keːli ₁	gətəhən ₂	gedehen ₂	gedyː ₂	geːs ₂	gets ₂	gesn ₂	gitis ₂ , xiwɮʃij ₋₁	ketysy ₂	ketesən ₂	kele ₁
pássaro	dəgi ₋₁	eiwan ₁	ʃubuːn ₁	ʃuwuː ₁	ʃuvuː ₁	ʃuwuː ₁	ʃowun ₁	ʃuwu ₁	ʃuβuː ₁	ʃuːn ₁ , peltʃər ₂	bendzer ₂
fogo	gali ₁	gal ₁	gal ₁	gal ₁	gal ₁	gal ₁	gal ₁	gal ₁	qal ₁	qal ₁	χal ₁
caminho	terg-uːl ₁	qargöi ₂	χargi ₂ , zam ₋₁	zam ₋₁	dzam ₋₁	dzam ₋₁	xaː-lkə ₃	tsam ₋₁	tʃam ₋₁	mør ₄	mor ₄
sal	hataː ₁	dawhən ₂	dabhan ₂	dawuhuː ₂	daws ₂	daws ₂	dawsn ₂	tawsă ₂	taβusu ₂	taːpsən ₂	dabsun ₂
nadar	unpa-du ₁	umba ₋₁	t ^h amar ₋₂	umb ₋₁	sele ₋₃	umba ₋₁	us-təi ₋₄ , øːm ₋₅	siɮi ₋₃	usu-tʃ ^h i-la ₋₄	umpa ₋₁	mba ₋₁
água	ə̆sə ₁	oxon ₁	uhan ₁	uːha ₁	usn ₁	us ₁	usn ₁	ʊsə̆ ₁	usun ₁	q ^h usun ₁	sə ₁
vento	kei ^h n ₁	halkin ₂	halxin ₂	halxi ₂	salʃxin ₂	salkʃi ₂	salʃkn ₂	saɮxi ₂	k ^h iː ₁	k ^h iː ₁	ki ₁

Parte V. Família yumana (México, EUA)

(F) (8 pontos) Examinem a seguinte lista. Abaixo vocês podem ver uma árvore construída com base na mesma lista. Alguns dados (nomes de línguas e distâncias lexicostatísticas) estão faltando. Preencham as lacunas. Indiquem se a árvore é baseada em anotações manuais ou automáticas, bem como se ela foi gerada usando o algoritmo A ou B.

	mojave	cocopa	yavapai	tiipay (Jamul)	'iipay (Mesa Grande)
curto	wena=wen-a ₁	'xɬ=ʔut ₂	ʔkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
pássaro	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=ʔʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:₂
osso	n=a=s-ak ₁	'n=j=a:k ₁	ʔʃ=j=a:k-a ₁	'ak ₁	aq ₁
seco	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:='θo-β-a ₃	'k ^w ak ₄	kuk ^w a:j-p ₁
pescoço	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:=puk ₂
ver	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
rabo	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
dois	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
ano	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=ʔʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 pontos) Foram geradas algumas outras árvores para a família yumana, com as seguintes distâncias lexicostatísticas na raiz da árvore (as distâncias lexicostatísticas na margem esquerda de cada árvore):

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

Desenhem cada uma dessas árvores. Para cada uma das árvores, indiquem se ela é baseada em anotações manuais ou automáticas, bem como se ela foi gerada usando o algoritmo A ou B.

(H) (3 pontos) Duas distâncias listadas na tarefa (G) foram arredondadas para duas casas decimais: o valor 0,23 foi obtido por meio de arredondamento de 0,225. Qual outra distância foi arredondada e qual é seu valor exato?

(I) (4 pontos) Expliquem como são calculados os índices de estabilidade.

(J) (5 pontos) Expliquem como são calculadas as distâncias lexicostatísticas.

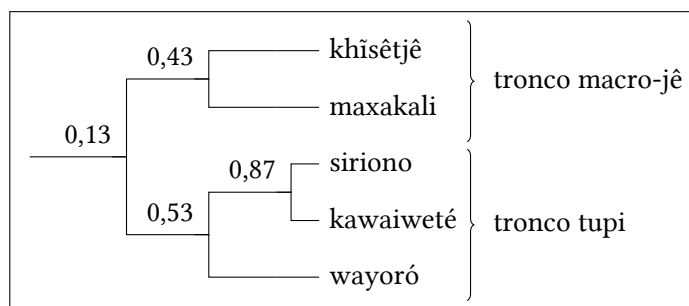
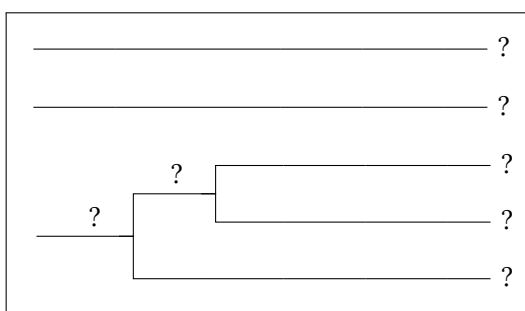
(K) (4 pontos) Expliquem a diferença entre os algoritmos A e B.

Parte VI. Tronco macro-jê e tronco tupi (Brasil, Bolívia)

(L) (28 pontos) Os troncos macro-jê e tupi são dois grandes agrupamentos linguísticos da América do Sul. Alguns linguistas acreditam que há uma relação distante entre eles. Examinem as seguintes listas.

	A	B	Γ	Δ	E
casca	e='e-ke	h ^w i='k ^h λ	kup='pe	mīβ̄m='t̄caj	= 'pe
barriga	'e=rje	= 't ^h igi	=ã'ñ	= 'tæj	=re'wεk
sangue	e='ruki	=ka' ⁿ b̄o	=d̄z=a'u	= 'hεβ̄p	=ru'i
queimar	= 'raī	=rɔ='k ^h λ̄ã	=po'k ^w a	mū=...='haβ̄p	=ra'pi
gordura	e='kira	= 't ^h wəmi	= 'd̄z=ap	= 'tuβ̄p	= 'kap
pé	'e=i	= 'h ^w aji	= 'β̄i	=po'ta	= 'pi
mão	'e=o	=nī'k ^h λa	= 'β̄o	= 'nīβ̄m	= 'pɔ
pesado	e='usi	=wi't ^h ī	=po'ti	=β̄p'təj	=pɔ'ij
fígado	'e=ja	= 'nba	=pi'a	=t̄caīβ̄pkī'nāj	=pi'ʔa
novo	e='jasu	= 'ndiwi	=pa'gop	= 'tiβ̄p	=pia'u
raiz	e='rao	=ja'ɾe	kup=kujop	mīβ̄m=nīβ̄m=t̄ca'tī	=ra'pɔ
pele	'e=i	= 'k ^h λ	= 'pe	= 't̄caj	= 'pit
rabo	e='rokōi	= 'nbi	=d̄z=ɔ'k ^w aj	=nā:='kiβ̄p	= 'raj
branco	'e=ʃī	=ja'k ^h a	=d̄zi'ra	=β̄p'dou	= 'sīŋ
asa	e='heo	=ja'ɾa	=pe'o	=nī'māu	=pe'pɔ, =ji'wa

Abaixo vocês podem ver duas árvores construídas com base nas mesmas listas. Alguns dados (nomes de línguas e distâncias lexicostatísticas) estão faltando. Preencham as lacunas. Para cada uma das árvores, indiquem se ela é baseada em anotações manuais ou automáticas, bem como se ela foi gerada usando o algoritmo A ou B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ As anotações manuais e os índices de estabilidade nesta tarefa foram omitidos intencionalmente.

(M) (10 pontos) Os procedimentos automatizados baseados nas classes de Dolgopolsky podem dar resultados errados. Neste exemplo, o procedimento automatizado detecta mais semelhanças entre o siriono e uma determinada língua macro-jê (khîsêtjê) do que entre o siriono e outras línguas do tronco tupi. Proponham e descrevam *brevemente* um procedimento modificado que resultaria em uma classificação correta se aplicado às listas macro-jê e tupi acima.

⚠ Esta tarefa será corrigida somente em caso de empate entre times mais bem classificados.

Os autores da questão agradecem a Alejandra Vidal, Maria Konoshenko, Iliá Gruntov e Jamthô Suyá por terem respondido perguntas sobre línguas específicas. —*Andrey Nikulin, Milena Vêneva*

Editores: Ivan Derjanski (editor técnico), Hugh Dobbs, Stanislav Guriévitch, Boris Iomdin, Liam McKnight, Andrey Nikulin (editor-chefe), Aleksejs Peguševs, Jan Petr, Alexander Piperski, Maria Rubinstein, Milena Vêneva, Elysia Warner.

Texto em português: Andrey Nikulin.

Boa prova!