

Vigésimoprimer Olimpiada Internacional de Lingüística

Brasilia (Brasil), 23–31 de julio de 2024

Problema del torneo de grupos

La lexicoestadística es un grupo de métodos creados para evaluar qué tan distante es la relación entre determinadas lenguas con base en su vocabulario. Normalmente esos métodos se aplican a largas listas de palabras anotadas por expertos, quienes señalan si, según su juicio, un determinado par de palabras proviene de la misma fuente. Sin embargo, a veces los lingüistas aplican métodos lexicoestadísticos a listas de palabras anotadas mediante procedimientos automatizados. Uno de esos procedimientos se basa en el concepto de *clases consonánticas*, introducido por Aharon Dolgopolsky, un lingüista soviético e israelí, en 1964.

P.	p b ɸ β f v	K.	k g x ɣ q ɕ χ ɰ	Y.	j ç (en el inicio de la raíz)	M.	m ɱ
T.	t d ɖ θ ð ʈ ɖ	R.	r ɾ ʀ ɹ l ʎ ʟ ʟʃ	W.	w ɰ (en el inicio de la raíz)	N.	n ɲ ɳ ɳ
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ					Q.	ʈ ɖ
H.	h ʕ ɦ ʕ ʔ h ɦ ʔ, vocales y j ç w ɰ (excepto en el comienzo de la raíz)						

Clases consonánticas de Dolgopolsky

A continuación, encontrarán fragmentos de listas de palabras de distintas familias lingüísticas del mundo. Las anotaciones se representan mediante índices numéricos. Con base en esas listas, se construyeron árboles de las respectivas familias lingüísticas usando dos versiones simplificadas del llamado algoritmo *StarlingNj*, y se atribuyó un *índice de estabilidad* a cada palabra. Los árboles y los índices de estabilidad en la parte superior se basan en listas anotadas manualmente, mientras que los de la parte inferior se basan en listas anotadas automáticamente. En cada caso, hay dos árboles para cada lista, construidas con versiones distintas del algoritmo: el algoritmo A y el algoritmo B. Nótese que en algunos casos hay múltiples árboles posibles que corresponden a una lista de palabras; en esos casos se eligió un árbol aleatoriamente. A cada nudo en cada árbol se atribuye una distancia lexicoestadística. Mientras mayor es la distancia, más cercana la relación entre las lenguas. De este modo, el término más exacto sería «distancia lexicoestadística invertida» y no «distancia lexicoestadística». Por simplicidad, en este problema utilizamos el término «distancia lexicoestadística».

Tanto los índices de estabilidad como las distancias lexicoestadísticas fueron redondeadas a dos cifras decimales. Si la tercera cifra decimal es inferior a 5, el valor se redondea por defecto; en el caso contrario, se redondea por exceso. Por ejemplo, 2,836 se redondea a 2,84, 0,705 a 0,71, 0,703 a 0,70. El redondeamiento se aplica únicamente a los valores mostrados a lectores humanos. Es decir, la máquina «ve» los valores no redondeados mientras aplica los algoritmos.

Nótese que algunas palabras son préstamos conocidos o probables de otras lenguas. Por ejemplo, la palabra **jok:i** ‘sal’ de la lengua kadiweo es un préstamo del guaraní (**juki**), mientras que la palabra **ʔa:nʔ** ‘año’ del ’iipai (Mesa Grande) es un préstamo del castellano (**‘año**).

A veces se listan múltiples sinónimos, separados por comas, que corresponden a un solo concepto. Un ejemplo es el concepto ‘pie’ en la lengua vejoj.

En los datos que siguen, todos los prefijos se separan mediante el símbolo «=», mientras que los sufijos se separan mediante el símbolo «-». Algunas palabras jamás ocurren sin prefijos. Éstas comienzan con el símbolo «=».

Los datos aparecen transcritos usando el Alfabeto Fonético Internacional. ^ˈ = acento primario, _ˌ = acento secundario (más débil que el primario), ː = sonido largo, ˘ = sonido muy corto, \overline{XY} = X y Y se

pronuncian como un único sonido, \acute{o} = tono alto, \grave{o} = tono bajo, \hat{o} = tono descendente, $^{\circ}$ = consonante preglotalizada (precedida por un corto bloqueo del flujo de aire en la glotis), $^{\circ}$ = consonante eyectiva (acompañada de un corto bloqueo del flujo de aire en la glotis), $^{\circ}$ = sonido sordo, $\tilde{\circ}$ = sonido nasalizado (pronunciado por la nariz), $^{\circ}$ = laringización (sonido bajo, crujiente), $^{\circ}$ señala un flujo de aire a través de la nariz antes de la consonante, $^{\circ h}$ = consonante aspirada (pronunciada con un soplo de aire), $^{\circ w}$ = consonante labializada (pronunciada con los labios redondeados), $^{\circ j}$ = consonante palatalizada (una parte de la lengua se acerca al palato duro). $\alpha, \text{æ}, \text{ɛ}, \text{ɪ}, \text{ɨ}, \text{ɔ}, \text{ʊ}, \text{u}, \text{ə}, \text{ʌ}, \text{ɒ}, \text{ɘ}, \text{y}, \text{ø}, \text{ɤ}$ son vocales. Otros símbolos especiales denotan consonantes.

⚠ El conocimiento de las lenguas mencionadas en este problema no da ninguna ventaja para su solución.

Parte I. La familia guaycurú (Argentina, Brasil, Paraguay)

	toba (oriental)	pilagá	mocoví (chaqueño)	kadiweo
nube	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
fuego	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
pez	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-ḏʒegi ₃
cabeza	=qajk ₁	=qajk ₁	=qaik ₁	=ak:ilo ₂
matar	=alawat ₁	=a ^h la:t ₁	=alawat ₁	=el:owadi ₁
luna	ʔawoxojk ₁	ʔa ^h woʎojk ₁	ʃirajyo ₂	ep:enaj ₃
nariz	=mik ₁	=mik ₁	=mik ₁	=m:iq:o ₁
sal	towe ₁	ol'ʔek ₂	ʔwe ₁	jok:i ₋₁
pedra	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
lengua	=atʃ-ak ^h at ₁	=a ^h tʃ-aʎat ₁	=oʔley-ak ^h an-ak ^h at ₂	=ok:el:i ₃

	algoritmo A	algoritmo B	
manual	<p>distancia lexicostadística</p>		índices de estabilidad: nube 0,50 fuego 0,50 pez 0,50 cabeza 0,75 matar 1,00 luna 0,50 nariz 1,00 sal 0,67 piedra 0,75 lengua 0,50
automática			índices de estabilidad: nube 0,50 fuego 0,50 pez 0,75 cabeza 0,75 matar 1,00 luna 0,50 nariz 1,00 sal 0,25 piedra 0,75 lengua 0,50

Parte II. La familia nubia (Egipto, Sudán)

	dongolau	kenuzi	dilling	kadaru	debri	birgid
matar	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
luna	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
agua	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
dar	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
bueno	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
viento	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
pelo	'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	ɟer₁	dwalleli₂	ɟer-i₁	ne:r-i₁
sol	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	algoritmo A	algoritmo B	
manual			Índices de estabilidad: matar 0,50 luna 0,83 agua 1,00 dar 1,00 bueno 0,50 viento 0,50 pelo 0,83 barriga 0,83 dormir 0,83 sol 0,50
automática			Índices de estabilidad: matar 0,33 luna 0,50 agua 0,50 dar 0,67 bueno 0,50 viento 0,50 pelo 0,83 barriga 1,00 dormir 0,50 sol 0,50

- (A) (2 puntos) La consonante **ɟ** se pronuncia como la *r* del francés, con la parte trasera de la lengua. ¿A qué clase de Dolgopolsky pertenece y cómo lo descubrieron?
- (B) (2 puntos) El árbol nubio en la parte superior izquierda es apenas uno de dos árboles posibles para esta combinación del algoritmo y del tipo de las anotaciones. Dibujen el otro árbol posible.
- (C) (2 puntos) El árbol nubio en la parte inferior izquierda es apenas uno de dos árboles posibles para esta combinación del algoritmo y del tipo de las anotaciones. Dibujen el otro árbol posible.
- (D) (2 puntos) La distancia lexicoestadística 0,49, atribuida a la raíz del árbol nubio en la parte superior derecha, fue redondeada a dos cifras decimales, como otras distancias en este problema. ¿Cuál es la distancia exacta?

Parte III. La familia mataguaya (Argentina, Bolivia, Paraguay)

	wichí (ber-mejeño abajeño)	wichí (Rivadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaçle (shichaam lhavos)	nivaçle (chisham-nee lhavos)	maká
fuego	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʔh ^w at ₂	ʔe ^w it'e ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
pez	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
pie	=patʃu ₁	=qolɔ ₂	=patʃo ₁ , =kala ₂	=pa:k'ɔʔ ₁	=ʔsat ₃	=ka'laʔ ₂	=φoʔ ₄	=φoʔ ₄	=f'iʔ ₅
agua	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔi'n'at ₁	ʔa'nat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
dar	=ʔwenɔ-u ₁	=wenɔ-u ₁	=ʔwenɔ-o ₁	=ʔwenɔ-oʔ ₁	=ʔwehn-aʔm ₂	=ʔhajʔ ₃ , =ʔwen ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
bueno	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejk'un-ej ₂
viento	ʔinwok ^w ₁	ʔinwɔk ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔhlahwuʔ ₄	ʔhlahwu ^w ʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	t'unik'i ₆
árbol	haʔlo ₁	halɔ ₁	haʔla ₁	haʔlaʔ ₁	ʔa'laʔ ₁	ʔa'la-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
pelo	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔateʔtʃ ₃	=jeʔs ₄	=ʔewkux-its ₅
matar	=lon ₁	=lɔn ₁	=lan ₁	=la:nɔ ₁	=ʔlaʔan ₁	=ʔlan ₁	=kla ^w n ₁	=kla ^w n ₁	=lan ₁

	algoritmo A	algoritmo B	
manual	<p> wichí (bermejeño abajeño) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaçle (shichaam lhavos) nivaçle (chishamnee lhavos) maká </p>	<p> wichí (bermejeño abajeño) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaçle (shichaam lhavos) nivaçle (chishamnee lhavos) maká </p>	índices de estabilidad: fuego 0,78 pez 1,00 pie 0,33 agua 0,78 dar 0,44 bueno 0,89 viento 0,33 árbol 0,78 pelo 0,67 matar 1,00
automática	<p> wichí (bermejeño abajeño) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaçle (shichaam lhavos) nivaçle (chishamnee lhavos) maká </p>	<p> wichí (bermejeño abajeño) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaçle (shichaam lhavos) nivaçle (chishamnee lhavos) maká </p>	índices de estabilidad: fuego 0,78 pez 0,44 pie 0,33 agua 0,56 dar 0,67 bueno 0,89 viento 0,22 árbol 0,67 pelo 0,67 matar 1,00

Parte IV. La familia mongólica (República Popular China, Mongolia, Rusia)

(E) (10 puntos) Examen la lista siguiente. Calculen los índices de estabilidad que corresponden a las anotaciones manuales y automáticas.

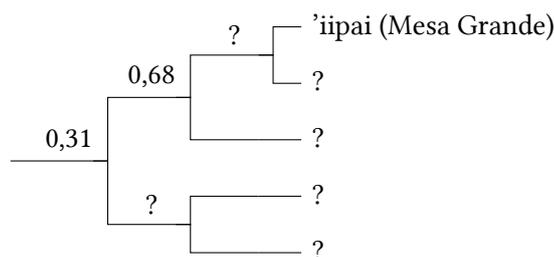
Para hacer su tarea más fácil, nosotros ya calculamos ambos índices para el concepto ‘todos’. En orden aleatorio, son: 0,36 y 0,40.

	daur (jai-lar)	jamnigan (manchú)	buriato (jori)	bargut nuevo	eleuto	joshut	calmuco	jalja	ordós	shira yugur	bonán
todos	hɔ:₁	bolt₂	boxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamăġ-₁	pyyyte₄, xamukᵃ-₁	tʃᵃuq₅	hanə-₂
corteza	hails₁	qalihon₁	χoltəhən₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəłtᵃšs₂	turusu₃	χalsən₁	arasun₄
barriga	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwłjij-₁	ketysy₂	ketesən₂	kele₁
pájaro	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendzer₂
fuego	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
camino	terg-u:l₁	qargöi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-lkə₃	tsam-₁	tjam-₁	mør₄	mor₄
sal	hata:₁	dawhən₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsun₂
nadar	unpa-du₁	umba-₁	tᵃamar-₂	umb-₁	sele-₃	umba-₁	us-tᵃi-₄, ø:m-₅	siłi-₃	usu-tʃᵃi-la-₄	umpa-₁	mba-₁
agua	əṣə₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ʊṣṣ₁	usun₁	qᵃusun₁	sə₁
viento	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkʃi₂	salʃkn₂	sałxı₂	kᵃi:₁	kᵃi:₁	ki₁

Parte V. La familia yumana (México, EE.UU.)

(F) (8 puntos) Examinen la lista siguiente. Abajo ustedes pueden observar un árbol construido con base en la misma lista. Algunos datos (nombres de lenguas y distancias lexicoestadísticas) fueron omitidos. Llenen las lagunas. Indiquen si el árbol se basa en anotaciones manuales o automáticas, así como si fue generado mediante el algoritmo A o B.

	mojave	cucapá	yavapai	tiipai (Jamul)	'iipai (Mesa Grande)
corto	wena=wen-a ₁	'xɬ=ʔut ₂	ʔkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
pájaro	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=ʔʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:₂
hueso	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
cuello	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
cola	i:=ʔar ₁	'ʃ=juʎ₂	'β=hé ₃	ʃə='juʎ₂	xə=juʎ₂
dos	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
año	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=ʔ ^h ur-a ₃	mat-'wam ₂	ʔa:n'₁



(G) (20 puntos) Se generaron algunos otros árboles para la familia yumana, con las siguientes distancias lexicoestadísticas en la raíz (las cuales se encuentran en el margen izquierdo de cada árbol):

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

Dibujen cada uno de estos árboles. Para cada árbol indiquen si se basa en anotaciones manuales o automáticas, así como si fue generado mediante el algoritmo A o B.

(H) (3 puntos) En la tarea (G), dos distancias fueron redondeadas a dos cifras decimales: el valor 0,23 se obtuvo mediante redondeamiento de 0,225. ¿Cuál otra distancia fue redondeada y cuál es su valor exacto?

(I) (4 puntos) Expliquen cómo se calculan los índices de estabilidad.

(J) (5 puntos) Expliquen cómo se calculan las distancias lexicoestadísticas.

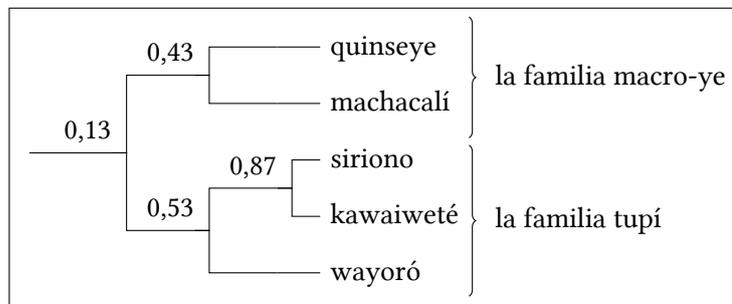
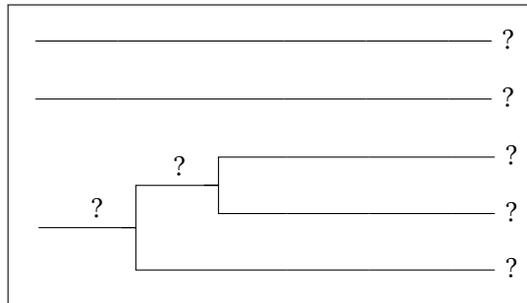
(K) (4 puntos) Expliquen la diferencia entre los algoritmos A y B.

Parte VI. La familia macro-ye y la familia tupí (Brasil, Bolivia)

(L) (28 puntos) Las familias macro-ye y tupí son dos familias lingüísticas importantes de Sudamérica. Algunos lingüistas creen que están lejanamente emparentadas. Examinen las listas siguientes.

	A	B	Γ	Δ	E
corteza	e='e-ke	h ^w i='k ^h Λ	kup='pε	mīβm=τεaj	= 'pε
barriga	'e=rje	=t ^h igi	=ã'ũn	= 'tæj	=rε'wεk
sangre	e='ruki	=ka' ⁿ bɾo	=d̄z=a'a	= 'hεβp	=ru'i
quemar	= 'raī	=rɔ='k ^h Λs̄	=po'k ^w a	mũ=...='haβp	=ra'pī
grasa	e='kira	=t ^h wəmi	= 'd̄z=ap	= 'tuβp	= 'kap
pie	'e=i	= 'h ^w aji	= 'βi	=pv'ta	= 'pī
mano	'e=o	=ɾi' ^k Λa	= 'βo	= 'ɾiβm	= 'pɔ
pesado	e='usi	=wi't ^h i	=po'ti	=βp'təj	=pɔ'ij
hígado	'e=ja	= 'nba	=pi'a	=τεiβpkī'nāj	=pī'ʔa
nuevo	e='jasu	= 'ndiwi	=pa'gop	= 'tiβp	=pia'u
raíz	e='rao	=ja'ɾe	kup=kujɔ'pε	mīβm=ɾiβm=τεa'tiə	=ra'pɔ
piel	'e=i	= 'k ^h Λ	= 'pε	= 'τεaj	= 'pit
cola	e='rokoī	= 'nbi	=d̄z=ɔ'k ^w aj	=nā:='kiβp	= 'raj
blanco	'e=fī	=ja'k ^h a	=d̄zi'ra	=βp'douɥ	= 'sīɾj
ala	e='heo	=ja'ɾa	=pε'o	=ɾi'māuɥ	=pε'pɔ, =ji'wa

Abajo ustedes pueden observar dos árboles construidos con base en las mismas listas. Algunos datos (nombres de lenguas y distancias lexicoestadísticas) fueron omitidos. Llenen las lagunas. Para cada árbol indiquen si se basa en anotaciones manuales o automáticas, así como si fue generado mediante el algoritmo A o B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ Las anotaciones manuales y los índices de estabilidad en esta tarea han sido omitidos de manera intencional.

(M) (10 puntos) Los procedimientos automáticos basados en las clases de Dolgopolsky pueden arrojar resultados incorrectos. En este ejemplo, el procedimiento automático detecta más semejanzas entre el siriono y una determinada lengua macro-ye (quinseye) que entre el siriono y las demás lenguas tupíes. Propongan y describan *brevemente* un procedimiento modificado que generaría una clasificación correcta de las lenguas macro-ye y tupíes a partir de las listas dadas.

⚠ Esta tarea se evaluará únicamente en caso de empate entre equipos con mayor puntuación.

Los autores del problema agradecen a Alejandra Vidal, María Konoshenko, Ilyá Gruntov y Jamthô Suyá por haber respondido preguntas sobre lenguas específicas. —*Andrey Nikulin, Milena Véneva*

Editores: Iván Derzhanski (editor técnico), Hugh Dobbs, Stanislav Gurévich, Borís Iomdin, Liam McKnight, Andrey Nikulin (editor jefe), Aleksejs Peguševs, Jan Petr, Alexánder Piperski, María Rubinstein, Milena Véneva, Elysia Warner.

Texto en castellano: Andrey Nikulin, Jenifer Vega Rodríguez.

¡Suerte!