

Twenty-first International Linguistics Olympiad

Brasília (Brazil), July 23–31, 2024

Team Contest Problem

Lexicostatistics is a group of methods designed to estimate how closely any languages are related to each other based on their vocabulary. These methods are normally applied to lengthy lists of words manually annotated by experts, who indicate whether any specific pair of words is believed to originate from the same source. Sometimes, however, linguists apply lexicostatistical methods to wordlists annotated by means of automated procedures. One such procedure is based on the concept of *consonant classes*, introduced by the Soviet–Israeli linguist Aharon Dolgopolsky in 1964.

P.	p b ɓ φ β f v	K.	k g x ɣ q ɠ χ w	Y.	j ç (root-initially)	M.	m ɱ
T.	t d ɗ θ ð ʈ ɖ	R.	r r̥ ɽ ɺ ɻ ʂ ʃ ʎ ʟ	W.	w ɯ (root-initially)	N.	n ɲ ɳ ŋ
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ ʑ					Q.	ʈ ɖ ʂ ʃ
H.	h ʕ ɦ ʕ̣ ʔ h ɦ ʔ, vowels, and j ç w ɯ (except root-initially)						

Dolgopolsky's consonant classes

Below you will find annotated fragments of wordlists of several language families of the world. The annotations are given with subscript digits. Based on these lists, language family trees have been constructed using two simplified versions of the so-called *StarlingNj* algorithm, and a *stability index* has been assigned to each word. The trees and stability indices on the top are based on manually annotated wordlists, and those on the bottom are based on lists that have been automatically annotated. There are two constructed trees for each wordlist, following two versions of the algorithm: Algorithm A and Algorithm B. Note that in some cases there are multiple possible trees corresponding to a wordlist; in such cases, only one tree was randomly chosen. Each node on each tree has a lexicostatistical distance assigned to it. The greater the distance, the closer the relationship between the languages. A more precise term would thus be “inverted lexicostatistical distance” rather than “lexicostatistical distance”. For simplicity's sake, we use the term “lexicostatistical distance” in this problem.

Both the stability indices and the lexicostatistical distances are rounded to two decimal places. If the third digit after the decimal point is smaller than 5, round down; otherwise, round up. For instance, 2.836 is rounded to 2.84, 0.705 is rounded to 0.71, and 0.703 is rounded to 0.70. The rounding applies only to the values shown to human readers. In other words, the computer that is running the algorithms “sees” the unrounded values.

Note that some words are known or suspected to have been borrowed from other languages. For example, the Kadiwéu word **jok:i** ‘salt’ is borrowed from Guaraní **juki**, and 'Iipay (Mesa Grande) **?a:nj** ‘year’ is borrowed from Spanish **'ano**.

In some cases multiple synonyms for a single meaning are given in the wordlists, separated by a comma. One example is ‘foot’ in Vejoz.

In the data below, all prefixes are separated by a “=” sign, and all suffixes are separated by a “-” sign. Some words are only ever used with prefixes. These start with a “=” sign.

The data are transcribed using the International Phonetic Alphabet. ¹ = primary stress, ₁ = secondary stress (weaker than the primary stress), ɔː = long sound, ɔ̃ = very short sound, X̂Y = X and Y are pronounced as one sound, ́ = high tone, ̀ = low tone, ˆ = falling tone, ʔ = preglottalized sound (preceded by a brief blocking of the flow of air in the throat), ʔ' = ejective sound (pronounced by briefly blocking the flow of air in the throat), ɸ = voiceless sound, ɸ̃ = nasalized sound (pronounced through

the nose), ɔ̥ = creaky voice (a low, scratchy sound), ⁿɔ̥ indicates some air flows through the nose before the consonant, ɔ^h = aspirated consonant (pronounced with a puff of air), ɔ^w = labialized consonant (pronounced with rounded lips), ɔ^j = palatalized sound (pronounced while part of the tongue is moved close to the hard palate). **ɑ, æ, ɛ, ɪ, ɨ, ɔ, ʊ, ʉ, ə, ʌ, ɒ, ɘ, ɤ, ɐ, ø** are vowels. Other special characters are consonants.

⚠ Knowledge of any of the languages mentioned in the problem does not give an advantage when solving the problem.

Part I. Guaicuruan family (Argentina, Brazil, Paraguay)

	Toba (Eastern)	Pilagá	Mocoví (Chaco)	Kadiwéu
cloud	l=ʔok ₁	'lo=ʔok ₁	nawexelek ₂	lol:adi ₃
fire	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
fish	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-ɖʒegi ₃
head	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
to kill	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
moon	ʔawoxojk ₁	ʔa'woʃojk ₁	ʃirajxo ₂	ep:enaj ₃
nose	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
salt	towe ₁	ol'ʒek ₂	ʔwe ₁	jok:i- ₁
stone	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
tongue	=atʃ-aʂat ₁	=a'tʃ-aʂat ₁	=oʔley-aʂan-aʂat ₂	=ok:el:i ₃

	Algorithm A	Algorithm B	
manual	<p>lexicostatistical distance</p>		Stability indices: cloud 0.50 fire 0.50 fish 0.50 head 0.75 to kill 1.00 moon 0.50 nose 1.00 salt 0.67 stone 0.75 tongue 0.50
automated			Stability indices: cloud 0.50 fire 0.50 fish 0.75 head 0.75 to kill 1.00 moon 0.50 nose 1.00 salt 0.25 stone 0.75 tongue 0.50

Part II. Nubian family (Egypt, Sudan)

	Dongolawi	Kenuzi	Dilling	Kadaru	Debri	Birgid
to kill	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
moon	u'n-at-t₁	an-at-ti₁	nən-ti₁	nən-tu₁	nən-to₁	ma:l₂
water	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
to give	'tir₁	tir₁	ti₁	ti₁	ti₁	te:n₁
good	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
wind	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
hair	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
belly	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
to sleep	'nɛ:r₁	ne:r₁	ɟer₁	dwallɛli₂	ɟer-i₁	ne:r-i₁
sun	'masil₁	masil₁	ɛɟ₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	Algorithm A	Algorithm B	
manual			Stability indices: to kill 0.50 moon 0.83 water 1.00 to give 1.00 good 0.50 wind 0.50 hair 0.83 belly 0.83 to sleep 0.83 sun 0.50
automated			Stability indices: to kill 0.33 moon 0.50 water 0.50 to give 0.67 good 0.50 wind 0.50 hair 0.83 belly 1.00 to sleep 0.50 sun 0.50

- (A) (2 points) The consonant **ɟ** is pronounced like French *r*, at the back of the tongue. Which Dolgopolsky class does it belong to, and how did you find that out?
- (B) (2 points) The Nubian tree on the top left is just one of two possible trees for this combination of algorithm and annotation type. Draw the other possible tree.
- (C) (2 points) The Nubian tree on the bottom left is just one of two possible trees for this combination of algorithm and annotation type. Draw the other possible tree.
- (D) (2 points) The lexicostatistical distance 0.49 (assigned to the root of the Nubian tree on the top right) has been rounded to two decimal places, like some other distances in this problem. What is the exact distance?

Part III. Mataguayan family (Argentina, Bolivia, Paraguay)

	Wichí (Lower Bermejeño)	Wichí (Riva- davia)	Vejoz	'Weenhayek	Iyojwa'aja'	Manjui	Nivaçle (Shichaam Lhavos)	Nivaçle (Chisham- nee Lhavos)	Maká
fire	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʔhwat ₂	ʔeite ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
fish	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
foot	=patʃu ₁	=qolɔ ₂	=patʃo ₁ , =kala ₂	=pa:kʔoʔ ₁	=ʔsat ₃	=kaʔlaʔ ₂	=φoʔ ₄	=φoʔ ₄	=fʔiʔ ₅
water	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
to give	=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
good	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
wind	ʔinwok ^w ₁	ʔinwɔk ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔhlahwuʔ ₄	ʔhlahwuuʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
tree	haʔlo ₁	halɔ ₁	haʔla ₁	haʔlaʔ ₁	ʔaʔlaʔ ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
hair	=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
to kill	=lon ₁	=lɔn ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=klan ₁	=klan ₁	=lan ₁

	Algorithm A	Algorithm B	
manual			Stability indices: fire 0.78 fish 1.00 foot 0.33 water 0.78 to give 0.44 good 0.89 wind 0.33 tree 0.78 hair 0.67 to kill 1.00
automated			Stability indices: fire 0.78 fish 0.44 foot 0.33 water 0.56 to give 0.67 good 0.89 wind 0.22 tree 0.67 hair 0.67 to kill 1.00

Part IV. Mongolic family (People's Republic of China, Mongolia, Russia)

(E) (10 points) Examine the following wordlist. Calculate the stability indices corresponding to both the manual and the automated annotations.

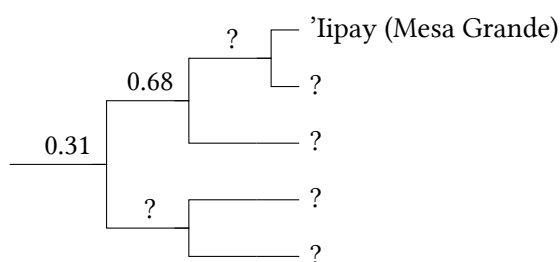
In order to help you, we have already calculated the two stability indices for the word 'all'. In random order, these are: 0.36 and 0.40.

	Dagur (Hailar)	Khamnigan (Manchu)	Buryat (Khorl)	New Bargut	Ööld	Khoshut	Kalmyk	Khalkha	Ordos	Shira Yugur	Bonan
all	hɔ:₁	bolt₂	bɔxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamăġ-₁	pyyyte₄, xamukᵃ-₁	tʃᵃuq₅	hanə-₂
bark	hails₁	qalihon₁	χoltōhōn₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəłtᵃᵃs₂	turusu₃	χalsən₁	arasun₄
belly	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwłjij-₁	ketysy₂	ketesən₂	kele₁
bird	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃor₂	bendzer₂
fire	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
road	terg-u:l₁	qargōi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-lkə₃	tsam-₁	tjam-₁	mør₄	mor₄
salt	hata:₁	dawhōn₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsun₂
to swim	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
water	ɔsə₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ʊsö₁	usun₁	qᵃusun₁	sə₁
wind	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkʃi₂	salʃkn₂	sałxı₂	kᵃi:₁	kᵃi:₁	ki₁

Part V. Yuman family (Mexico, United States of America)

(F) (8 points) Examine the following wordlist. Below you can see a tree that was built based on the same list. Some data (language names and lexicostatistical distances) are missing. Fill in the gaps. Specify if the tree is manual or automated, as well as if it was generated using Algorithm A or B.

	Mojave	Cocopa	Yavapai	Tiipay (Jamul)	'Iipay (Mesa Grande)
short	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuɲ ₁	mə=put-k ₃
bird	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ='ʃa ₂	ʔa:=ʃa:₂
bone	ɲ=a=s=ak ₁	'ɲ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
dry	i=ro:-v-k ₁	's=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
meat	k ^{wi} :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
neck	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:=puk ₂	i:=puk ₂
to see	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
tail	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
two	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
year	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 points) Some other trees have been generated for Yuman, with the following lexicostatistical distances at the root of the tree (the lexicostatistical distances on the very left of each tree):

1. 0.20
2. 0.23
3. 0.24

Draw each of these trees. For each of the trees, specify if it is manual or automated, as well as if it was generated using Algorithm A or B.

(H) (3 points) Two of the distances listed in Assignment (G) have been rounded to two decimal places: 0.23 has been rounded from 0.225. Which other distance has been rounded, and what is its precise value?

(I) (4 points) Explain how the stability indices are calculated.

(J) (5 points) Explain how the lexicostatistical distances are calculated.

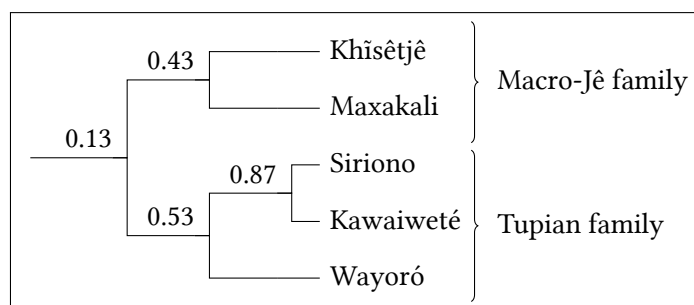
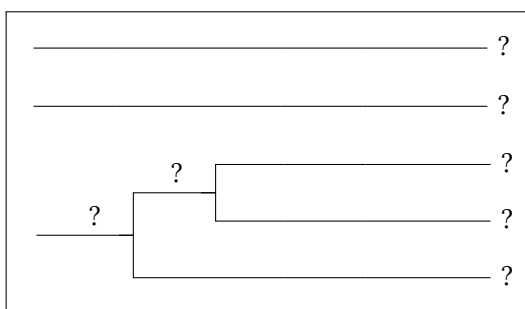
(K) (4 points) Explain the difference between Algorithms A and B.

Part VI. Macro-Jê family and Tupian family (Brazil, Bolivia)

(L) (28 points) Macro-Jê and Tupian are two major language families of South America. Some linguists believe them to be distantly related. Examine the following wordlists.

	A	B	Γ	Δ	E
bark	e='e-ke	h ^w ĩ='k ^h Λ	kup='pe	mĩβm='tɛaj	= 'pe
belly	'e=rje	= 't ^h igi	=ã'ũn	= 'təj	=rɛ'wek
blood	e='ruki	=ka' ⁿ bɾo	=d̄z=a'a	= 'hɛβp	=ru'i
to burn	= 'raĩ	=rɔ='k ^h ɹã	=po'k ^w a	mũ=...='haβp	=ra'pi
fat	e='kira	= 't ^h wəmi	=d̄z=ap	= 'tuβp	= 'kap
foot	'e=i	= 'h ^w aji	= 'βi	=pɔ'ta	= 'pi
hand	'e=o	=ɲĩ'k ^h ɹa	= 'βo	= 'ɲĩβm	= 'pɔ
heavy	e='usi	=wi't ^h ĩ	=po'ti	=βp'təj	=pɔ'ij
liver	'e=ja	= 'nba	=pi'a	=tɛiβpkĩ'nāj	=pi'ʔa
new	e='jasu	= 'ndiwi	=pa'gop	= 'tiβp	=pia'u
root	e='rao	=ja'ɾe	kup=kujɔ'pe	mĩβm=ɲĩβm=tɛa'tiə	=ra'pɔ
skin	'e=i	= 'k ^h Λ	= 'pe	= 'tɛaj	= 'pit
tail	e='rokoĩ	= 'nbi	=d̄z=ɔ'k ^w aj	=nã:'kiβp	= 'raj
white	'e=ʃĩ	=ja'k ^h a	=d̄zi'ra	=βp'dou	= 'sĩɲ
wing	e='heo	=ja'ɾa	=pe'o	=ɲĩ'māu	=pe'pɔ, =ji'wa

Below you can see two trees that were built based on the same lists. Some data (language names and lexicostatistical distances) are missing. Fill in the gaps. For each of the trees, specify if it is manual or automated, as well as if it was generated using Algorithm A or B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ The manual annotations and the stability indices were intentionally omitted in this assignment.

(M) (10 points) Automated procedures based on Dolgopolsky classes may yield incorrect results. In this example, the automated procedure detects more similarities between Siriono and a certain Macro-Jê language (Khîsêtjê) than between Siriono and other Tupian languages. Propose a modified automated procedure that would yield a correct classification if applied to the Macro-Jê and Tupian wordlists above, and describe it *briefly*.

⚠ This assignment will be graded only in the event of a tie between top-scoring teams.

The authors thank Alejandra Vidal, Maria Konoshenko, Ilya Gruntov, and Jamthô Suyá for answering their questions on specific languages.
—*Andrey Nikulin, Milena Veneva*

Editors: Ivan Derzhanski (technical editor), Hugh Dobbs, Stanislav Gurevich, Boris Iomdin, Liam McKnight, Andrey Nikulin (editor-in-chief), Aleksejs Peguševs, Jan Petr, Alexander Piperski, Maria Rubinstein, Milena Veneva, Elysia Warner.

English text: Andrey Nikulin, Milena Veneva.

Good luck!