en(B)

Twenty-first International Linguistics Olympiad

Brasília (Brazil), 23–31 July 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 ɸ β f v	K.	k g x ɣ q ɢ χ ɯ	Y.	j ç (root-initially)	M.	m ɱ
T.	t d ɗ θ ð ʈ ɖ	R.	r ɾ ɽ ɹ l ɬ ɮ ɭ ʎ ɬ	W.	w ʍ (root-initially)	N.	n ɳ ɲ ŋ
S.	s z ʃ ʒ ʂ ʐ ɕ ʑ c ɟ					Q.	t͡ɬ d͡ɮ
H.	ħ ʕ ʜ ʢ ʡ h ɦ ʔ, vowels, and j ç w ʍ (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 **jokːi** 'salt' is borrowed from Guaraní **jukɨ**, and 'Iipay (Mesa Grande) **ʔaːnʲ** 'year' is borrowed from Spanish **ˈaɲo**.

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.

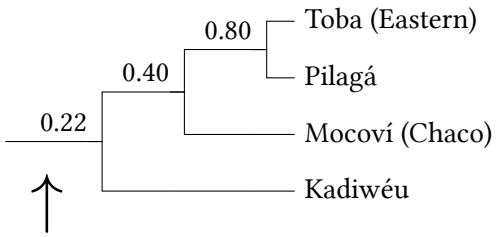
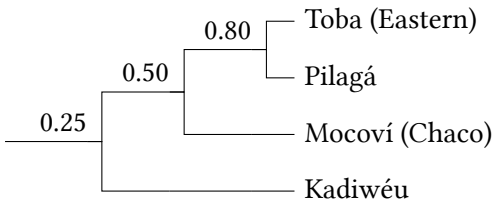
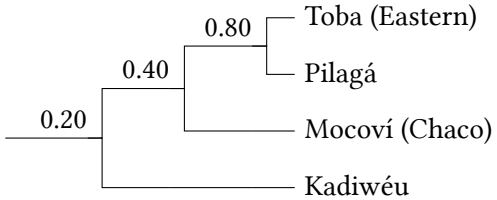
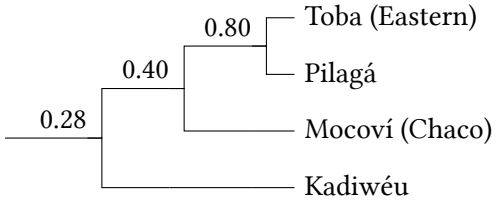
The data are transcribed using the International Phonetic Alphabet. ˈ = primary stress, ˌ = secondary stress (weaker than the primary stress), ◌ː = long sound, ◌̆ = very short sound, X͡Y = X and Y are pronounced as one sound, ◌́ = high tone, ◌̀ = low tone, ◌̂ = falling tone, ◌ˀ = preglottalised sound (preceded by a brief blocking of the flow of air in the throat), ◌' = ejective sound (pronounced by briefly blocking the flow of air in the throat), ◌̥ = voiceless sound, ◌̃ = nasalised sound (pronounced through

the nose), ◌̰ = creaky voice (a low, scratchy sound), ⁿ◌ indicates some air flows through the nose before the consonant, ◌ʰ = aspirated consonant (pronounced with a puff of air), ◌ʷ = labialised consonant (pronounced with rounded lips), ◌ʲ = palatalised sound (pronounced while part of the tongue is moved close to the hard palate). **ɑ**, **æ**, **ɛ**, **ɪ**, **i̇**, **ɔ**, **ʊ**, **ʉ**, **ə**, **ʌ**, **ɒ**, **ɘ**, **ɣ**, **ɵ**, **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.

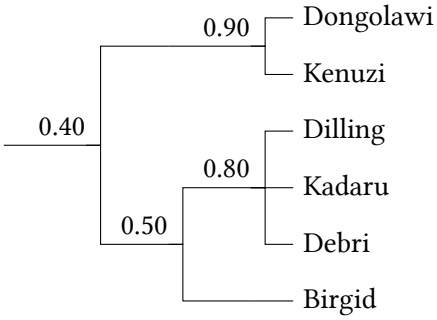
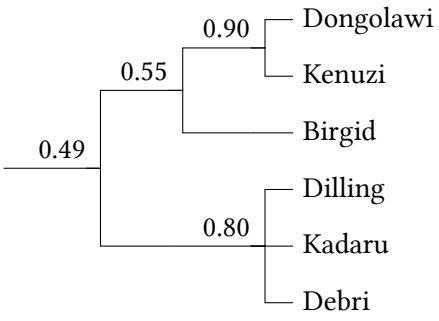
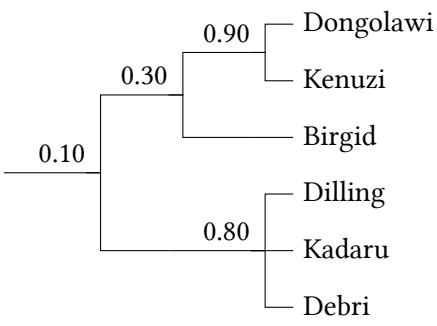
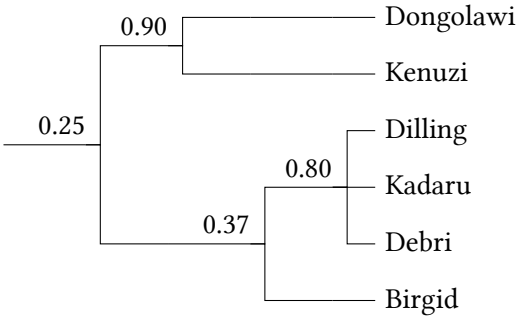
Part I. Guaicuruan family (Argentina, Brazil, Paraguay)

	Toba (Eastern)	Pilagá	Mocoví (Chaco)	Kadiwéu
cloud	l=ʔok$_1$	ˈlo=ʔok$_1$	naweɣelek$_2$	lol:adi$_3$
fire	nodek$_1$	ˈd=oleʔ$_2$	noɾek$_1$	n=ol:edi$_2$
fish	ɳjaq$_1$	ˈnijaq$_1$	naʌ̈in$_2$	nij:oɢo-d͡ʒegi$_3$
head	=qajk$_1$	=ˈqajk$_1$	=qaik$_1$	=ak:ilo$_2$
to kill	=alawat$_1$	=aˈla:t$_1$	=alawat$_1$	=el:owadi$_1$
moon	ʔawoʁojk$_1$	ʔaˈwoʃojk$_1$	ʃɪrajɣo$_2$	ep:enaj$_3$
nose	=mik$_1$	=ˈmik$_1$	=mik$_1$	=m:iq:o$_1$
salt	towe$_1$	olˈɣek$_2$	ʔwe$_1$	jok:i$_{-1}$
stone	qaʔ$_1$	ˈqaʔ$_1$	qaʔ$_1$	wet:iɢa$_2$
tongue	=at͡ʃ-aʁat$_1$	=aˈt͡ʃ-aʃat$_1$	=oʔleɣ-aʁan-aʁat$_2$	=ok:el:i$_3$

	Algorithm A	Algorithm B	
manual	 lexicostatistical distance		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	ˈbɛː$_1$	beː$_1$	hur$_2$	wur-i$_2$	wur-i$_2$	filaːl-e$_1$
moon	uˈn-at-t$_1$	an-at-ti$_1$	nɔn-ti$_1$	nɔn-tu$_1$	nɔn-to$_1$	maːl$_2$
water	ˈɛss$_1$	essi$_1$	ɔti$_1$	ɔto$_1$	ɔtu$_1$	eɟi$_1$
to give	ˈtɪr$_1$	tir$_1$	ti$_1$	ti$_1$	ti$_1$	teː-n$_1$
good	ˈsɛrɛː$_1$	sereː$_1$	ken$_2$	kɛn$_2$	kɛŋ$_2$	azze-n$_3$
wind	ˈturug$_1$	turug$_1$	irʃ-i$_2$	irʃ-o$_2$	irʃ-o$_2$	kurr-i$_3$
hair	ˈdɪl-tɪ$_1$	siːr$_2$	tel-ti$_1$	til-tu$_1$	til-tu$_1$	ur=dill-e$_1$
belly	ˈtuː$_1$	tuː$_1$	te-te$_2$	to$_1$	to$_1$	tuː$_1$
to sleep	ˈnɛːr$_1$	neːr$_1$	ɟer$_1$	dwallɛli$_2$	ɟer-i$_1$	neːr-i$_1$
sun	ˈmasɪl$_1$	masɪl$_1$	ɛɟ$_2$	aɟu$_2$	eŋgal-to$_3$	ʔiːzi$_2$

	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 **ʁ** 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?

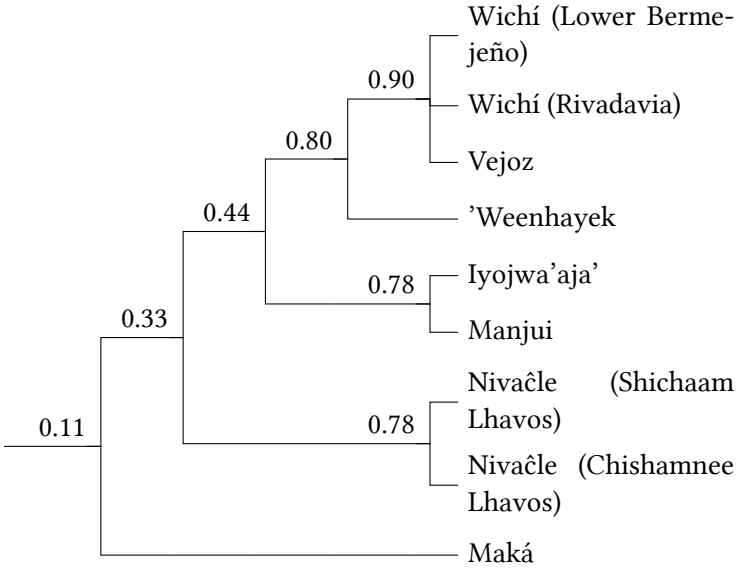
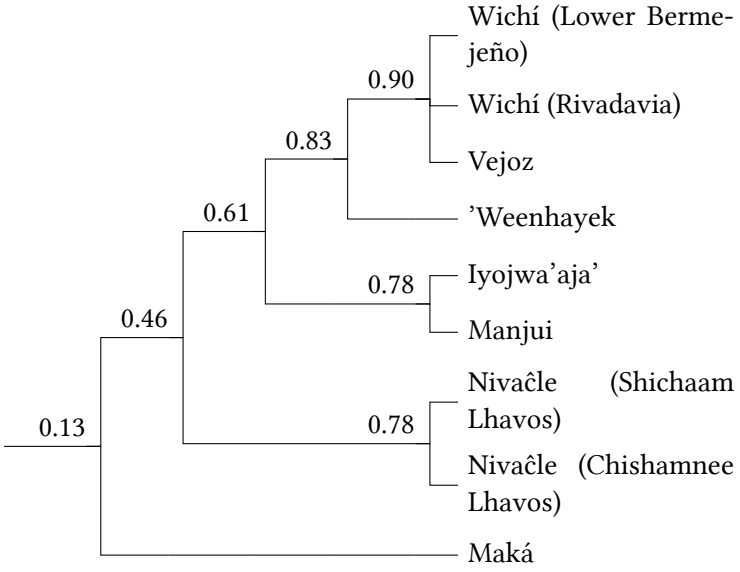
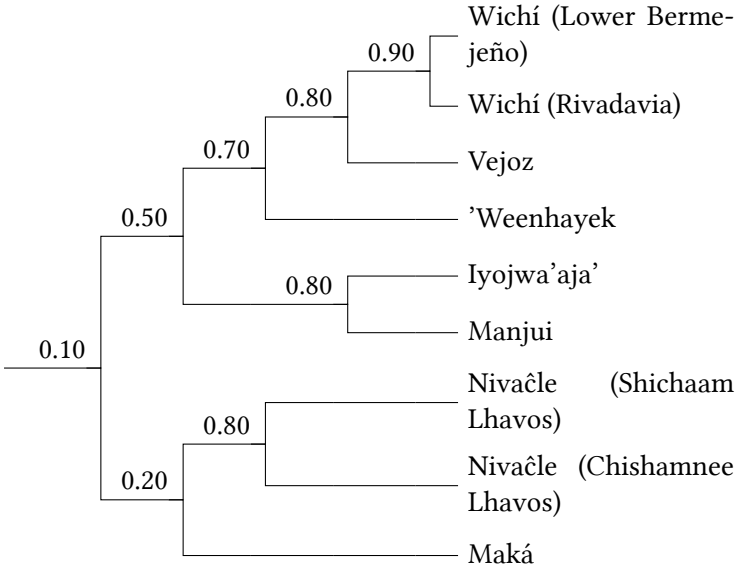
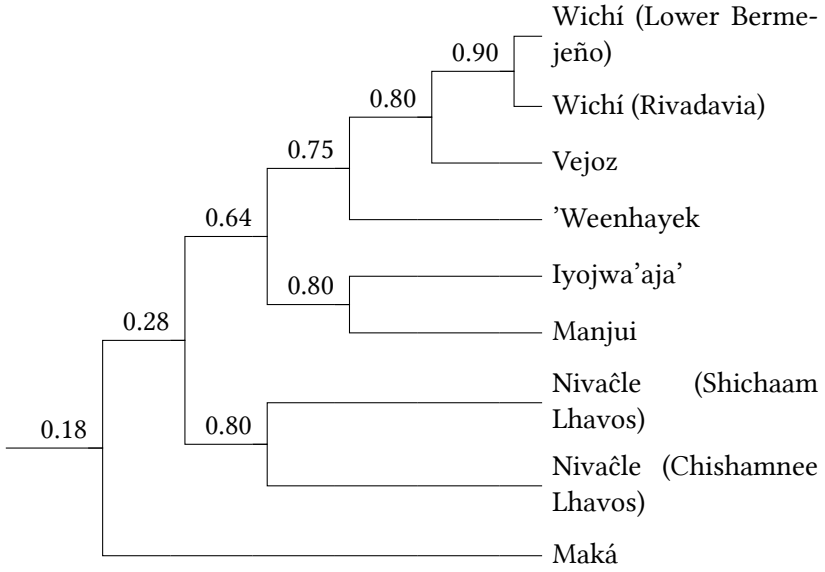
(B) **(2 points)** The Nubian tree on the top left is just one of two possible trees for this combination of algorithm and annotation type. Draw the other possible tree.

(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	ʾWeenhayek	Iyojwaʾajaʾ	Manjui	Nivaĉle (Shichaam Lhavos)	Nivaĉle (Chisham-nee Lhavos)	Maká
fire	ʔitoχ$_1$	ʔitɔx$_1$	ʔitɑh$_1$	ʔiːtɑχ$_1$	ˈh͡wat$_2$	ˈʔe͡itʲe$_1$	ʔitɑx$_1$	ʔitɑx$_1$	feʔt$_2$
fish	ʔwahat$_1$	wahat$_1$	wahat$_1$	ʔwaːhat$_1$	siˈʔjus$_{-1}$	ʃiˈʔjus$_{-1}$	saxet͡ʃ$_{-1}$	saxet͡ʃ$_{-1}$	seheth͡s$_{-1}$
foot	=pat͡ʃu$_1$	=qɔlɔ$_2$	=pat͡ʃo$_1$, =kɑla$_2$	=paːkʲʾoʔ$_1$	=ˈsat$_3$	=kaˈʔlaʔ$_2$	=ɸoʔ$_4$	=ɸoʔ$_4$	=fʾiʔ$_5$
water	ʔinot$_1$	ʔinɔt$_1$	wah$_2$	ʔinɑːt$_1$	ʔiˈnʲat$_1$	ʔaˈʔnat$_1$	jinɑʔt$_1$	jinɑʔt$_1$	iweliʔ$_3$
to give	=ʔweɲ̥-u$_1$	=weɲ̥-u$_1$	=ʔweɲ̥-o$_1$	=ʔweɲ̥-oʔ$_1$	=ˈwɛh͡n-aʔm$_2$	=ˈhajʔ$_3$, =ˈwɛn$_2$	=xut$_4$	=xut-ej$_4$	tis-ix$_5$
good	ʔis$_1$	ʔis$_1$	ʔis$_1$	ʔis$_1$	ˈʔes$_1$	ˈʔe͡is$_1$	ʔis$_1$	ʔis$_1$	t=ejkʾun-ej$_2$
wind	ʔinwokʷ$_1$	ʔinwɔk$_1$	ʔihwokʷ$_1$	=jaːɬ$_2$, =xʷoxʷ$_3$	ˈh͡lah͡wuʔ$_4$	ˈh͡lah͡wuuʔ$_4$	ɬaβiʔm$_5$	ɬaβiʔm$_5$	tʾunikʾi$_6$
tree	haʔlo$_1$	halɔ$_1$	haʔlɑ$_1$	haʔlɑʔ$_1$	ʔaˈʔlaʔ$_1$	ʔaˈʔla-k$_1$	ʔaʔkxi-juk$_2$	jiʔk͡lɑʔ$_1$	naxka-k$_3$
hair	=ʔwule-j$_1$	=wule-j$_1$	=ʔwole-j$_1$	=ʔwoːle-ç$_1$, hiːlenaχ$_2$	=ˈʔwole$_1$	=ˈʔwole-j$_1$	=ʃateʔt͡ʃ$_3$	=jeʔs$_4$	=ʔewkux-it͡s$_5$
to kill	=lon$_1$	=lɔn$_1$	=lɑn$_1$	=lɑːn̥$_1$	=ˈlaʔan$_1$	=ˈlan$_1$	=k͡lɑn$_1$	=k͡lɑn$_1$	=lan$_1$

<table>
<tr><th></th><th>Algorithm A</th><th>Algorithm B</th><th></th></tr>
<tr>
<td>manual</td>
<td> 
</td>
<td> 
</td>
<td>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</td>
</tr>
<tr>
<td>automated</td>
<td> 
</td>
<td> 
</td>
<td>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</td>
</tr>
</table>

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.

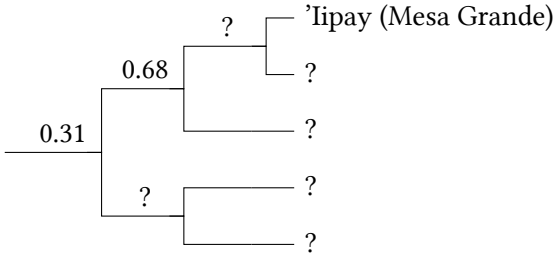
In order to help you, we have already calculated the two stability indices for the word 'all'. In random order, these are: 0.36 and 0.40.

	Dagur (Hailar)	Khamnigan (Manchu)	Buryat (Khori)	New Bargut	Ööld	Khoshut	Kalmyk	Khalkha	Ordos	Shira Yugur	Bonan
all	hɔː$_{1}$	bʊlt$_{2}$	bʊxɨː$_{3}$	bygd$_{4}$	t͡sug$_{5}$	lug$_{5}$	t͡suk$_{5}$, xamak$_{-1}$	pux$_{3}$, pugt$_{4}$, xamăɢ$_{-1}$	pyɣyte$_{4}$, xamuk^{h}$_{-1}$	t͡ʃ^{h}uq$_{5}$	hanə$_{-2}$
bark	hai̯l͡s$_{1}$	qalihʊn$_{1}$	χɔltəhɔn$_{2}$	xaʲhuː$_{1}$	xolts$_{2}$	xalis$_{1}$	dursn$_{3}$	xɔɮt^{h}ʊ̆s$_{2}$	turusu$_{3}$	χalsən$_{1}$	arasun$_{4}$
belly	keːli$_{1}$	gʊtəhʉn$_{2}$	gedehen$_{2}$	gedyː$_{2}$	geːs$_{2}$	gets$_{2}$	gesn$_{2}$	gitĭs$_{2}$, xiwɮĭj$_{-1}$	ketysy$_{2}$	ketesən$_{2}$	kele$_{1}$
bird	dəgi$_{-1}$	ɕiwan$_{1}$	ʃubuːn$_{1}$	ʃuwuː$_{1}$	ʃuvuː$_{1}$	ʃuwuː$_{1}$	ʃowun$_{1}$	ʃuwu$_{1}$	ʃuβuː$_{1}$	ʃuːn$_{1}$, pelt͡ʃɔr$_{2}$	bend͡ʑer$_{2}$
fire	gali$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	qal$_{1}$	qal$_{1}$	χal$_{1}$
road	terg-uːl$_{1}$	qarɢʊ̈i̯$_{2}$	χarɢɨ$_{2}$, zam$_{-1}$	zam$_{-1}$	d͡zam$_{-1}$	d͡zam$_{-1}$	xaː-lʁə$_{3}$	t͡sam$_{-1}$	t͡ʃam$_{-1}$	mør$_{4}$	mor$_{4}$
salt	hataː$_{1}$	dawhʊn$_{2}$	dabhan$_{2}$	dawuhuː$_{2}$	daws$_{2}$	daws$_{2}$	dawsn$_{2}$	tawsă$_{2}$	taβusu$_{2}$	taːpsən$_{2}$	dabsuŋ$_{2}$
to swim	unpa-du$_{1}$	ʊmba-$_{1}$	t^{h}amar-$_{2}$	umb-$_{1}$	sele-$_{3}$	umba-$_{1}$	us-t͡ɕi-$_{4}$, øːm-$_{5}$	siɮĭ-$_{3}$	usu-t͡ʃ^{h}i-la-$_{4}$	umpa-$_{1}$	mba-$_{1}$
water	ɔsɔ$_{1}$	ʊxʊn$_{1}$	uhan$_{1}$	uːha$_{1}$	usn$_{1}$	us$_{1}$	usn$_{1}$	ʊsŏ$_{1}$	usun$_{1}$	q^{h}usun$_{1}$	sə$_{1}$
wind	kei̯n$_{1}$	halkin$_{2}$	halxin$_{2}$	halxi$_{2}$	salʲxin$_{2}$	salkʲi$_{2}$	salʲkn$_{2}$	saɮxĭ$_{2}$	k^{h}iː$_{1}$	k^{h}iː$_{1}$	ki$_{1}$

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	ˈxʌ́=ʔut$_2$	ˈt͡ʃkr=ot-i_2	ləˈʔuɲ$_1$	mə=put-k_3
bird	ʔit͡ʃ=i=jer$_1$	ˈʃa$_2$	ˈʔ=t͡ʃ=sa$_2$	aʔ=ˈʃa$_2$	ʔaː=ʂaː$_2$
bone	ɲ=a=s=ak$_1$	ˈɲ=j=aːk$_1$	ˈt͡ʃ=j=aːk-a_1	ˈak$_1$	aq$_1$
dry	i=roː-v-k$_1$	ˈʂ=ʔar$_2$	ˈru-β-i_1	ˈs=ʔaːj$_3$	saːj$_3$
meat	kʷiːkʷay$_1$	ʔi=ˈmaːt͡ʃ$_2$	ˌkʷeː=ˈθo-β-a_3	ˈkʷak$_4$	kukʷaːj-p$_1$
neck	maʌ́aqe$_1$	ˈm=puk$_2$	ˈmlqí$_1$	iː=ˈpuk$_2$	iː=puk$_2$
to see	i=juː-k$_1$	ˈwiː$_2$	ˈʔuː$_1$	ˈwiːw$_2$	ə=wuːw$_2$
tail	iː=ʔar$_1$	ˈʃ=juʌ́$_2$	ˈβ=hé$_3$	ʃəˈjuɬ$_2$	xə=juɬ$_2$
two	havik-k$_1$	ˈx=wak$_1$	ˈh͡wâk-i_1	xəˈwak$_1$	xə=wak$_1$
year	huːðe$_1$	ˌmat-ˈkaːm$_2$	ˈʔ=t͡ʃʰur-a_3	mat-ˈwam$_2$	ʔaːnʲ$_{-1}$



(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.

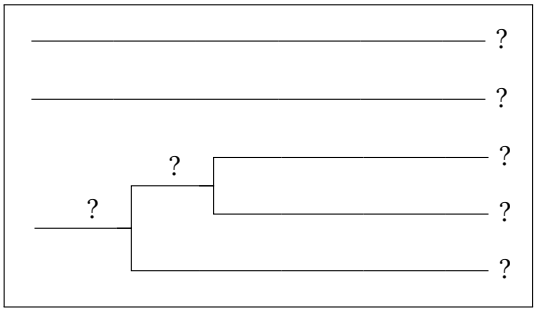
(K) **(4 points)** Explain the difference between Algorithms A and B.

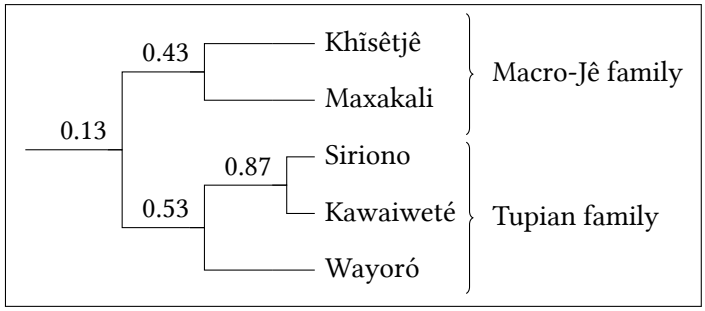
Part VI. Macro-Jê family and Tupian family (Brazil, Bolivia)

(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=ˈe-ke	hʷ̃ĩ=ˈkʰʌ	kʉp=ˈpɛ	mĩβ͡m=ˈt͡ɕaj	=ˈpɛ
belly	ˈe=ɾje	=ˈtʲʰigi	=ã̰ˈṵ̃n	=ˈtæj	=ɾɛˈwɛk
blood	e=ˈɾuki	=kaˈⁿbɻo	=d͡ʒ=aˈʉ	=ˈhɛβ͡p	=ɾuˈɨ
to burn	=ˈɾa͡i	=ɾɔ=ˈkʰɹɜ̃	=poˈkʷa	mũ=...=ˈha͡βp	=ɾaˈpɨ
fat	e=ˈkiɾa	=ˈtʲʰwɜmɨ	=ˈd͡ʒ=ap	=ˈtu͡βp	=ˈkap
foot	ˈe=i	=ˈhʷaji	=ˈβi	=pɒˈta	=ˈpɨ
hand	ˈe=o	=ɲɨ̃ˈkʰɹa	=ˈβo	=ˈɲĩβ͡m	=ˈpɔ
heavy	e=ˈusi	=wɨˈtʲʰĩ	=poˈti	=β͡pˈtəj	=pɔˈɨj
liver	ˈe=ja	=ˈⁿba	=piˈa	=t͡ɕɨβ͡pkɨ̃ˈnãj	=pɨˈʔa
new	e=ˈjasu	=ˈⁿdɨwɨ	=paˈgop	=ˈtɨβ͡p	=pɨaˈu
root	e=ˈɾao	=jaˈʈe	kʉp=kʉjoˈpɛ	mĩβ͡m=ɲĩβ͡m=t͡ɕaˈtɨ͡ə	=ɾaˈpɔ
skin	ˈe=ɨ	=ˈkʰʌ	=ˈpɛ	=ˈt͡ɕaj	=ˈpit
tail	e=ˈɾoko͡i	=ˈⁿbɨ	=d͡ʒ=oˈkʷaj	=nãː=ˈkɨβ͡p	=ˈɾaj
white	ˈe=ʃɨ̃	=jaˈkʰa	=d͡ʒiˈɾa	=β͡pˈdouɻ	=ˈsĩŋ
wing	e=ˈheo	=jaˈʈa	=pɛˈo	=ɲĩˈmãɯ̃ɻ	=pɛˈpɔ, =jɨˈwa

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 (Khĩsêtjê) 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, Eimear 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!