zh

第二十一届国际语言学奥林匹克竞赛

巴西·巴西利亚·2024年7月23 — 31日

团队赛题目

词汇统计法是一组根据词汇的相似程度来估算语言之间亲缘关系远近的方法。这类方法通常需要一个由专家手动标注过的冗长的单词表，标注的内容是单词表中的哪些单词被认为是同源的。不过，语言学家有时也将这类方法用在通过特定的流程自动生成标注的单词表上。其中一种流程基于*辅音类*这一概念，由俄裔以色利籍语言学家阿哈龙·多尔戈波尔斯基于1964年引入。

P.	p b ɓ ɸ β f v	K.	k g x ɣ q ɢ χ ɯ	Y.	j ç（在词根的开头）	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（在词根的开头）	N.	n ɳ ɲ ŋ
S.	s z ʃ ʒ ʂ ʐ ɕ ʑ ɕ ʝ					Q.	t͡ɬ d͡ɮ

H. ħ ʕ ʜ ʢ ʰ h ɦ ʔ，所有的元音，以及不在词根开头的j ç w ʍ

多尔戈波尔斯基的辅音类

本题将展示若干个已标注的单词表，每个单词表对应着一个语系。标注通过下标的数字表示。基于这几个单词表，语言学家通过名为“*StarlingNJ*”的算法的两个简化版本画出了语言谱系树状图，并计算了每个单词的*稳定指数*。在每个树状图表格中，上方的树状图和稳定指数基于的是手动标注的单词表，而下方的树状图和稳定指数基于的是自动标注的单词表。每种标注方式又各自对应两个树状图，分别通过算法A和算法B得到。需要注意的是，在某些情况下，一个单词表有不止一个可能的树状图。如果遇到这种情况，本题会随机展示这些树状图中的一个。在树状图的每个节点上都标有一个词汇统计距离。该距离越大，语言之间的亲缘关系越近。因此，“词汇统计距离”更准确的说法应该是“反词汇统计距离”。为简洁起见，本题依然采用“词汇统计距离”这个术语。

稳定指数和词汇统计距离均四舍五入至小数点后两位。若小数点后第三位小于5，则向下取整，否则向上取整。例如：2.836四舍五入为2.84；0.705四舍五入为0.71；0.703四舍五入为0.70。只有向人类读者展示的数值会被四舍五入。换言之，运行算法的计算机“看到”的是四舍五入之前的数值。

请注意，某些单词已知或疑似是外来词。例如：卡迪维奥语单词**jokːi** ‘盐’来自瓜拉尼语**jukɨ**；伊派语（梅萨格兰德方言）单词**ʔaːnʲ** ‘年’来自西班牙语**ˈaɲo**。

在某些情况下，单词表对于单个含义会列出多个同义词，用逗号分隔。例如贝霍斯语中的‘脚’。

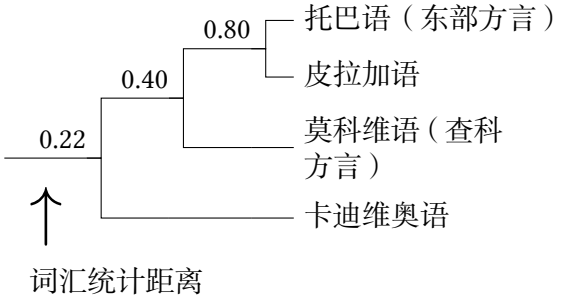
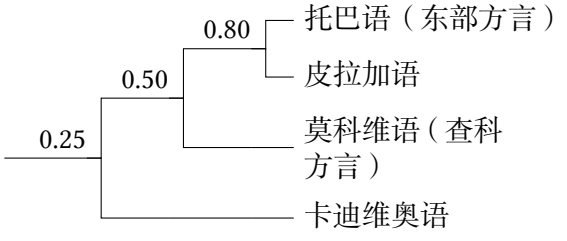
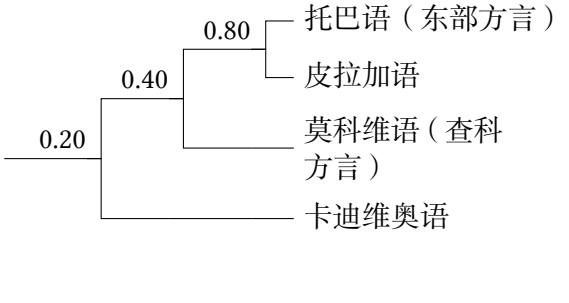
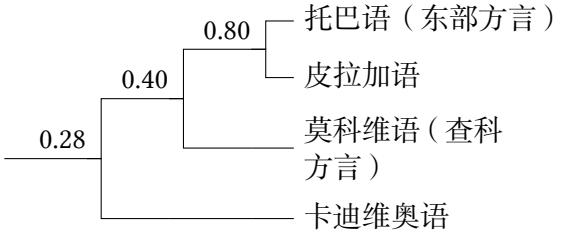
在以下语料中，所有前缀均用“=”符号分隔，所有后缀均用“-”符号分隔。有些单词必须加前缀。这些单词以“=”符号开头。

所有语料均采用国际音标转写。ˈ = 主重音、ˌ = 次重音（比主重音弱）、◌ː = 长音、◌̆ = 超短音、X͡Y = X和Y同时发音、◌́ = 高平调、◌̀ = 低平调、◌̂ = 降调、◌ˀ = 前声门化发音（发音前喉部短暂地阻塞气流）、◌’ = 外挤气音（发音时喉部短暂的阻塞气流）、◌̥ = 清音、◌̃ = 鼻音化发音（通过鼻腔发音）、◌̰ = 嘎裂声（低沉、沙哑的发音）、ⁿ◌表示鼻腔在辅音发音前有气流通过、◌ʰ = 送气辅音（发音时有一股气流喷出）、◌ʷ = 唇音化辅音（发音时圆唇）、◌ʲ = 硬颚化发音（发音时舌头的一部分靠近硬颚）。**ɑ**、**æ**、**ɛ**、**ɪ**、**ɨ**、**ɔ**、**ʊ**、**ʉ**、**ə**、**ʌ**、**ɒ**、**ɘ**、**y**、**ɵ**、**ø**为元音。其他特殊字符均为辅音。

⚠ 与本题涉及的语言有关的知识不会为解题带来任何优势。

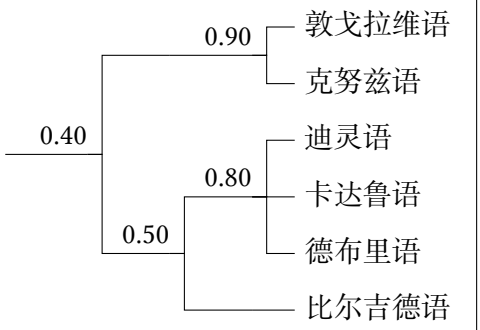
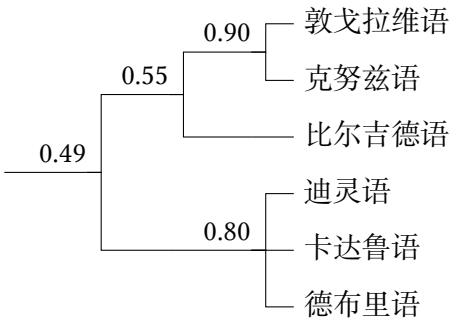
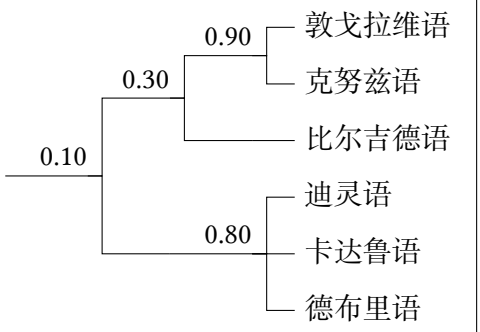
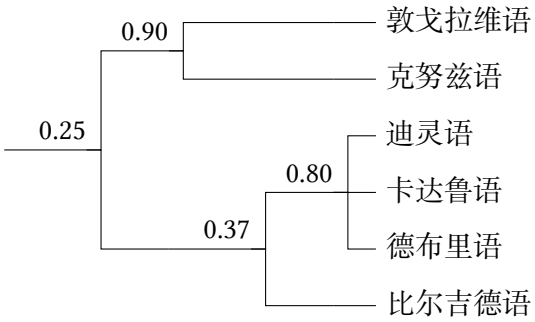
第壹部分.瓜伊库鲁语系（阿根廷、巴西、巴拉圭）

	托巴语（东部方言）	皮拉加语	莫科维语（查科方言）	卡迪维奥语
云	l=ʔok$_1$	ˈlo=ʔok$_1$	naweɣelek$_2$	lol:adi$_3$
火	nodek$_1$	ˈd=oleʔ$_2$	noɾek$_1$	n=ol:edi$_2$
鱼	njaq$_1$	ˈnijaq$_1$	naʌin$_2$	nij:oɢo-d͡ʒegi$_3$
头	=qajk$_1$	=ˈqajk$_1$	=qaik$_1$	=ak:ilo$_2$
杀	=alawat$_1$	=aˈla:t$_1$	=alawat$_1$	=el:owadi$_1$
月亮	ʔawoʁojk$_1$	ʔaˈwoʕojk$_1$	ʃiɾajɣo$_2$	ep:enaj$_3$
鼻子	=mik$_1$	=ˈmik$_1$	=mik$_1$	=m:iq:o$_1$
盐	towe$_1$	olˈɣek$_2$	ʔwe$_1$	jok:i$_{-1}$
石头	qaʔ$_1$	ˈqaʔ$_1$	qaʔ$_1$	wet:iɢa$_2$
舌头	=at͡ʃ-aʁat$_1$	=aˈt͡ʃ-aʕat$_1$	=oʔleɣ-aʁan-aʁat$_2$	=ok:el:i$_3$

<table>
<tr><th></th><th>算法A</th><th>算法B</th><th></th></tr>
<tr>
<td>手动</td>
<td> 
</td>
<td> 
</td>
<td>稳定指数：
云 0.50
火 0.50
鱼 0.50
头 0.75
杀 1.00
月亮 0.50
鼻子 1.00
盐 0.67
石头 0.75
舌头 0.50</td>
</tr>
<tr>
<td>自动</td>
<td> 
</td>
<td> 
</td>
<td>稳定指数：
云 0.50
火 0.50
鱼 0.75
头 0.75
杀 1.00
月亮 0.50
鼻子 1.00
盐 0.25
石头 0.75
舌头 0.50</td>
</tr>
</table>

第贰部分.努比亚语系（埃及、苏丹）

	敦戈拉维语	克努兹语	迪灵语	卡达鲁语	德布里语	比尔吉德语
杀	ˈbɛː$_1$	beː$_1$	hur$_2$	wur-i$_2$	wur-i$_2$	fila:l-e$_1$
月亮	uˈn-at-t$_1$	an-at-ti$_1$	nɔn-ti$_1$	nɔn-tu$_1$	nɔn-to$_1$	ma:l$_2$
水	ˈɛss$_1$	essi$_1$	ɔti$_1$	ɔto$_1$	ɔtu$_1$	eɟi$_1$
给	ˈtɪr$_1$	tir$_1$	ti$_1$	ti$_1$	ti$_1$	teː-n$_1$
好	ˈsɛrɛː$_1$	sereː$_1$	ken$_2$	kɛn$_2$	kɛŋ$_2$	azze-n$_3$
风	ˈturug$_1$	turug$_1$	irʃ-i$_2$	irʃ-o$_2$	irʃ-o$_2$	kurr-i$_3$
头发	ˈdɪl-tɪ$_1$	siːr$_2$	tel-ti$_1$	til-tu$_1$	til-tu$_1$	ur=dill-e$_1$
肚子	ˈtuː$_1$	tuː$_1$	te-te$_2$	to$_1$	to$_1$	tuː$_1$
睡觉	ˈnɛːr$_1$	neːr$_1$	ɟer$_1$	dwallɛli$_2$	ɟer-i$_1$	neːr-i$_1$
太阳	ˈmasɪl$_1$	masɪl$_1$	ɛɟ$_2$	aɟu$_2$	eŋgal-to$_3$	ʔiːzi$_2$

	算法A	算法B	
手动			稳定指数： 杀 0.50 月亮 0.83 水 1.00 给 1.00 好 0.50 风 0.50 头发 0.83 肚子 0.83 睡觉 0.83 太阳 0.50
自动			稳定指数： 杀 0.33 月亮 0.50 水 0.50 给 0.67 好 0.50 风 0.50 头发 0.83 肚子 1.00 睡觉 0.50 太阳 0.50

（A）（2分）辅音ʁ音似法语的r，发音部位在舌头后端。该辅音属于哪一个多尔戈波尔斯基类？你是如何得出这一结论的？

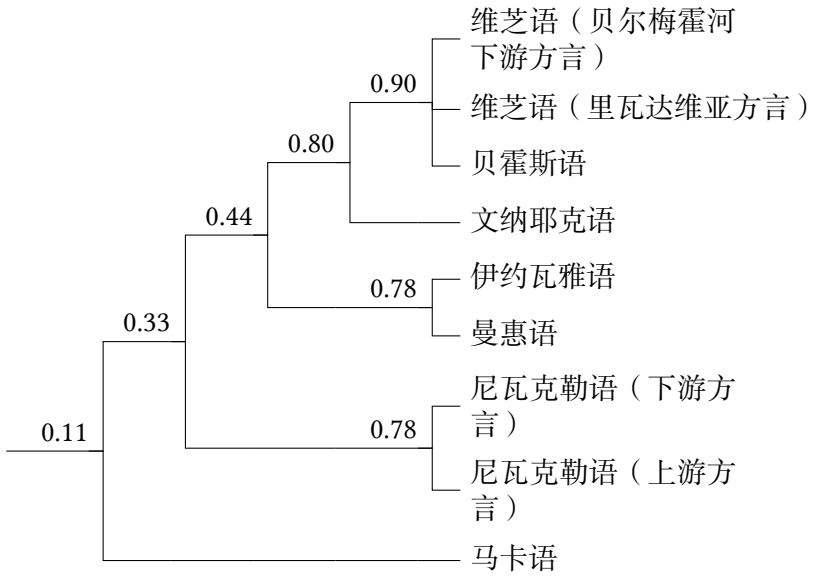
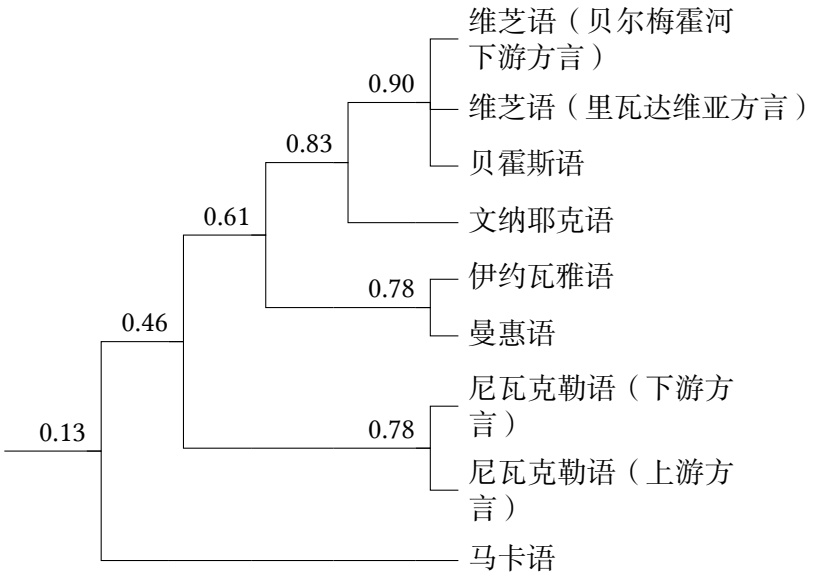
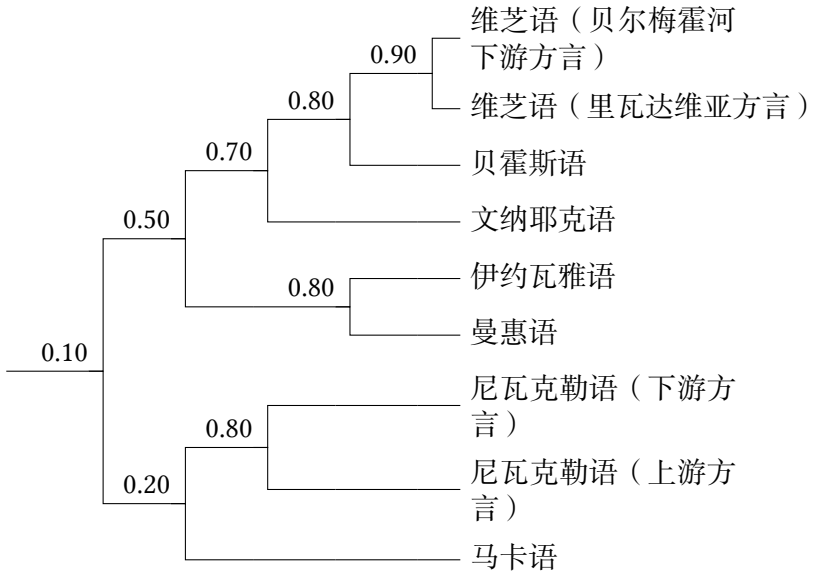
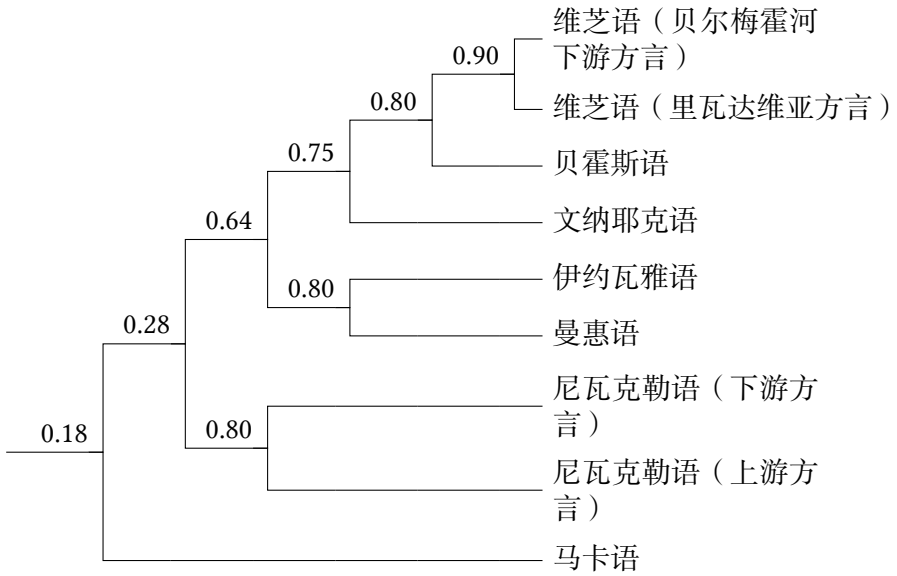
（B）（2分）左上方的努比亚语系树状图是该算法和标注方式下两种可能的树状图之一。画出另一个可能的树状图。

（C）（2分）左下方的努比亚语系树状图是该算法和标注方式下两种可能的树状图之一。画出另一个可能的树状图。

（D）（2分）正如本题中展示的其他词汇统计距离一样，右上方树状图根节点的距离0.49只保留了两位小数。该距离在四舍五入前的具体数值是多少？

第参部分.马塔瓜亚语系（阿根廷、玻利维亚、巴拉圭）

	维芝语（贝尔梅霍河下游方言）	维芝语（里瓦达维亚方言）	贝霍斯语	文纳耶克语	伊约瓦雅语	曼惠语	尼瓦克勒语（下游方言）	尼瓦克勒语（上游方言）	马卡语
火	ʔitoχ$_{1}$	ʔitɔx$_{1}$	ʔitah$_{1}$	ʔiːtɑχ$_{1}$	ˈh͡wat$_{2}$	ˈʔe͡itʲe$_{1}$	ʔitɑx$_{1}$	ʔitɑx$_{1}$	feʔt$_{2}$
鱼	ʔwahat$_{1}$	wahat$_{1}$	wahat$_{1}$	ʔwaːhat$_{1}$	siˈʔjus$_{-1}$	ʃiˈʔjus$_{-1}$	saxet͡ʃ$_{-1}$	saxet͡ʃ$_{-1}$	seheth͡s$_{-1}$
脚	=pat͡ʃu$_{1}$	=qɔlɔ$_{2}$	=pat͡ʃo$_{1}$, =kɑla$_{2}$	=paːkʲoʔ$_{1}$	=ˈsat$_{3}$	=kaˈʔlaʔ$_{2}$	=ɸoʔ$_{4}$	=ɸoʔ$_{4}$	=fʼiʔ$_{5}$
水	ʔinot$_{1}$	ʔinɔt$_{1}$	wah$_{2}$	ʔinɑːt$_{1}$	ʔiˈn̪at$_{1}$	ʔaˈʔnat$_{1}$	jinɑʔt$_{1}$	jinɑʔt$_{1}$	iweliʔ$_{3}$
给	=ʔweɲ̟-u$_{1}$	=weɲ̟-u$_{1}$	=ʔweɲ̟-o$_{1}$	=ʔweɲ̟-oʔ$_{1}$	=ˈwɛh͡n-aʔm$_{2}$	=ˈhajʔ$_{3}$, =ˈwɛn$_{2}$	=xut$_{4}$	=xut-ej$_{4}$	tis-ix$_{5}$
好	ʔis$_{1}$	ʔis$_{1}$	ʔis$_{1}$	ʔis$_{1}$	ˈʔes$_{1}$	ˈʔe͡is$_{1}$	ʔis$_{1}$	ʔis$_{1}$	t=ejkʼun-ej$_{2}$
风	ʔinwokʷ$_{1}$	ʔinwɔk$_{1}$	ʔihwokʷ$_{1}$	=jaːɬ$_{2}$, =xʷoxʷ$_{3}$	ˈh͡lah͡wuʔ$_{4}$	ˈh͡lah͡wuuʔ$_{4}$	ɬaβiʔm$_{5}$	ɬaβiʔm$_{5}$	tʼunikʼi$_{6}$
树	haʔlo$_{1}$	halɔ$_{1}$	haʔlɑ$_{1}$	haʔlɑʔ$_{1}$	ʔaˈʔlaʔ$_{1}$	ʔaˈʔla-k$_{1}$	ʔaʔkxi-juk$_{2}$	jiʔk͡lɑʔ$_{1}$	naxka-k$_{3}$
头发	=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$_{5}$
杀	=lon$_{1}$	=lɔn$_{1}$	=lɑn$_{1}$	=lɑːn̪$_{1}$	=ˈlaʔan$_{1}$	=ˈlan$_{1}$	=k͡lɑn$_{1}$	=k͡lɑn$_{1}$	=lan$_{1}$

	算法A	算法B	
手动			稳定指数： 火 0.78 鱼 1.00 脚 0.33 水 0.78 给 0.44 好 0.89 风 0.33 树 0.78 头发 0.67 杀 1.00
自动			稳定指数： 火 0.78 鱼 0.44 脚 0.33 水 0.56 给 0.67 好 0.89 风 0.22 树 0.67 头发 0.67 杀 1.00

第肆部分.蒙古语系（中国、蒙古、俄罗斯）

（E）（**10分**）仔细研究以下单词表。求与自动标注和手动标注对应的稳定指数。

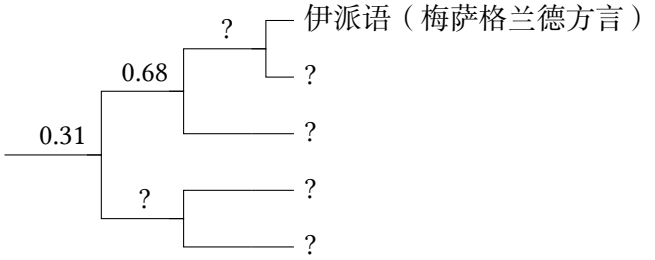
本题将单词‘所有’的两个稳定指数全部给出，以作提示。它们是：0.36、0.40（按随机顺序排列）。

	达斡尔语（海拉尔方言）	哈米尼干语（满洲方言）	布里亚特语（黜里方言）	新巴尔虎语	额鲁特语	和硕特语	卡尔梅克语	喀尔喀语	鄂尔多斯语	东部裕固语	保安语
所有	hɔː$_{1}$	bʊlt$_{2}$	bʊxiː$_{3}$	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$_{-2}$
树皮	hai̯l̩s$_{1}$	qalihʊn$_{1}$	χɔltəhɔn$_{2}$	xalʲhuː$_{1}$	xolts$_{2}$	xalis$_{1}$	dursn$_{3}$	xɔɮt^{h}ɔ̆s$_{2}$	turusu$_{3}$	χalsən$_{1}$	arasun$_{4}$
肚子	keːli$_{1}$	gətəhʉn$_{2}$	gedehen$_{2}$	gedyː$_{2}$	geːs$_{2}$	gets$_{2}$	gesn$_{2}$	gitĭs$_{2}$, xiwɮĭj$_{-1}$	ketysy$_{2}$	ketesən$_{2}$	kele$_{1}$
鸟	dəgi$_{-1}$	ɕiwan$_{1}$	ʃubuːn$_{1}$	ʃuwuː$_{1}$	ʃuvuː$_{1}$	ʃuwuː$_{1}$	ʃowun$_{1}$	ʃuwu$_{1}$	ʃuβuː$_{1}$	ʃuːn$_{1}$, pelt͡ʃɔr$_{2}$	bend͡zer$_{2}$
火	gali$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	ɢal$_{1}$	qal$_{1}$	qal$_{1}$	χal$_{1}$
路	terg-uːl$_{1}$	qarɢʊ̯i$_{2}$	χarɢi$_{2}$, zam$_{-1}$	zam$_{-1}$	d͡zam$_{-1}$	d͡zam$_{-1}$	xaː-lʁə$_{3}$	t͡sam$_{-1}$	t͡ʃam$_{-1}$	mør$_{4}$	mor$_{4}$
盐	hataː$_{1}$	dawhʊn$_{2}$	dabhan$_{2}$	dawuhuː$_{2}$	daws$_{2}$	daws$_{2}$	dawsn$_{2}$	tawsă$_{2}$	taβusu$_{2}$	taːpsən$_{2}$	dabsuŋ$_{2}$
游泳	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-$_{1}$
水	ɔsɔ$_{1}$	ʊxʊn$_{1}$	uhan$_{1}$	uːha$_{1}$	usn$_{1}$	us$_{1}$	usn$_{1}$	ʊsŏ$_{1}$	usun$_{1}$	q^{h}usun$_{1}$	sə$_{1}$
风	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$_{1}$

第伍部分.尤玛语系（墨西哥、美国）

（F）（**8分**）仔细研究以下单词表。下面是一个基于相同的单词表画出的树状图。其中缺少一些数据（语言名称和词汇统计距离）。填写其中的空缺。具体说明该树状图是基于手动标注还是自动标注，以及是用算法A还是算法B生成的。

	莫哈韦语	科科帕语	亚瓦派语	蒂派语（哈穆尔方言）	伊派语（梅萨格兰德方言）
短	wena=wen-a$_1$	ˈxʌ́=ʔut$_2$	ˈt͡ʃkr=ot-i$_2$	lə=ˈʔuɲ$_1$	mə=put-k$_3$
鸟	ʔit͡ʃ=i=jer$_1$	ˈʃa$_2$	ˈʔ=t͡ʃ=sa$_2$	aʔ=ˈʃa$_2$	ʔaː=ʂaː$_2$
骨头	ɲ=a=s=ak$_1$	ˈɲ=j=aːk$_1$	ˈt͡ʃ=j=aːk-a$_1$	ˈak$_1$	aq$_1$
干燥	i=roː-v-k$_1$	ˈʂ=ʔar$_2$	ˈru-β-i$_1$	ˈs=ʔaːj$_3$	saːj$_3$
肉	kʷiːkʷay$_1$	ʔi=ˈmaːt͡ʃ$_2$	ˌkʷeː=ˈθo-β-a$_3$	ˈkʷak$_4$	kukʷaːj-p$_1$
脖子	maʎaqe$_1$	ˈm=puk$_2$	ˈmlqíˀ$_1$	iː=ˈpuk$_2$	iː=puk$_2$
看到	i=juː-k$_1$	ˈwiː$_2$	ˈʔuː$_1$	ˈwiːw$_2$	ə=wuːw$_2$
尾巴	iː=ʔar$_1$	ˈʃ=juʂ$_2$	ˈβ=hé$_3$	ʃə=ˈjuɬ$_2$	xə=juɬ$_2$
二	havik-k$_1$	ˈx=wak$_1$	ˈh͡wâk-i$_1$	xə=ˈwak$_1$	xə=wak$_1$
年	huːðe$_1$	ˌmat-ˈkaːm$_2$	ˈʔ=t͡ʃʰur-a$_3$	mat-ˈwam$_2$	ʔaːnʲ$_{-1}$



（G）（**20分**）语言学家还另外生成了一些尤玛语系树状图。这些树状图根节点的词汇统计距离（即树状图上最左边的词汇统计距离）如下：

1. 0.20
2. 0.23
3. 0.24

画出这些树状图。具体说明每一个树状图是基于手动标注还是自动标注，以及是用算法A还是算法B生成的。

（H）（**3分**）小题（G）里词汇统计距离中的两个是在四舍五入至小数点后两位得到的：其中，0.23从0.225四舍五入得到。另一个在四舍五入后得到的距离是哪一个？该距离在四舍五入前的具体数值是多少？

（I）（**4分**）写出稳定指数的计算方法。

（J）（**5分**）写出词汇统计距离的计算方法。

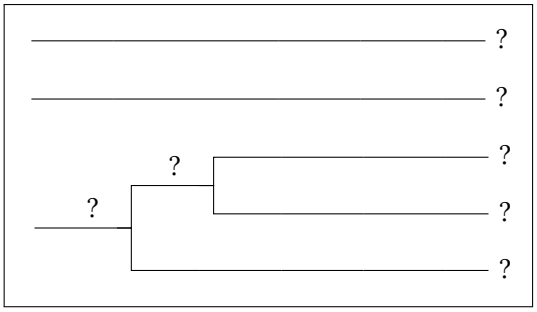
（K）（**4分**）写出算法A与B之间的区别。

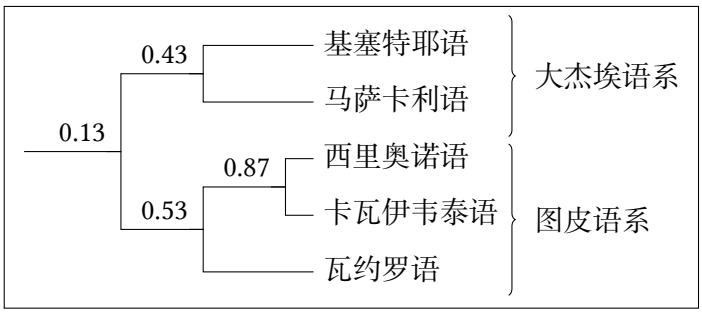
第陆部分.大杰埃语系、图皮语系（巴西、玻利维亚）

（L）（28分）大杰埃语系和图皮语系是南美洲的两个主要语系。一些语言学家认为这两个语系存在远亲关系。仔细研究以下单词表。

	A	B	Γ	Δ	E
树皮	e='e-ke	hʷĩ='kʰʌ	kʉp='pɛ	mĩβ͡m='t͡ɕɑj	='pɛ
肚子	'e=ɾje	='ʈʰigi	=ã'ũn	='tæj	=ɾɛ'wɛk
血	e='ɾuki	=ka'ⁿbɽo	=d͡ʒ=a'ʉ	='hɛβ͡p	=ɾu'ɨ
烧	='ɾai͡	=ɾɔ='kʰɹɜ̃	=po'kʷa	mũ=⋯='haβ͡p	=ɾa'pɨ
肥	e='kiɾa	='ʈʰwɜmɨ	='d͡ʒ=ap	='tuβ͡p	='kap
脚	'e=i	='hʷaji	='βi	=pɒ'ta	='pɨ
手	'e=o	=ɲĩ'kʰɹa	='βo	='ɲĩβ͡m	='pɔ
重	e='usi	=wɨ'ʈʰĩ	=po'ti	=β͡p'tɜj	=pɔ'ɨj
肝	'e=ja	='ⁿba	=pi'a	=t͡ɕɨβ͡pkĩ'nãj	=pɨ'ʔa
新	e='jasu	='ⁿdɨwɨ	=pa'gop	='tɨβ͡p	=pɨa'u
根	e='ɾao	=ja'ɽe	kʉp=kʉjo'pɛ	mĩβ͡m=ɲĩβ͡m=t͡ɕa'tiə͡	=ɾa'pɔ
皮肤	'e=ɨ	='kʰʌ	='pɛ	='t͡ɕɑj	='pit
尾巴	e='ɾoko͡i	='ⁿbɨ	=d͡ʒ=o'kʷaj	=nãː='kɨβ͡p	='ɾaj
白	'e=ʃɨ	=ja'kʰa	=d͡ʒi'ɾa	=β͡p'douɻ	='sĩŋ
翅膀	e='heo	=ja'ɽa	=pɛ'o	=ɲĩ'mãuɻ	=pɛ'pɔ, =jɨ'wa

下面是两个基于相同的单词表画出的树状图。其中缺少一些数据（语言名称和词汇统计距离）。填写其中的空缺。具体说明每一个树状图是基于手动标注还是自动标注，以及是用算法A还是算法B生成的。





A	B	Γ	Δ	E
?	?	?	?	?

⚠ 本小题省略了手动标注和稳定指数。

（M）（**10分**）基于多尔戈波尔斯基类的自动化流程可能产生错误的结果。在本题展示的例子中，该自动化流程的结果显示，西里奥诺语与大杰埃语系某个语言（基塞特耶语）的相似程度反而比西里奥诺语与图皮语系其他语言的相似程度高。提出一种优化的自动化流程，使得该流程能基于以上大杰埃和图皮语系单词表给出正确的分类结果，并对该流程进行简要的描述。

⚠ 本小题只有在高分团队平分时才计分。

作者感谢阿莱汉德拉·比达尔、玛丽娅·科诺申科、伊利亚·格伦托夫、亚姆托·苏亚回答有关特定语言的问题。 ——安德雷·尼库林、米莱娜·韦内娃

编者： 扬·彼得、戴谊凡（伊万·德尔然斯基）（技术编辑）、窦修（休·多布斯）、斯塔尼斯拉夫·古列维奇、玛丽娅·鲁宾斯坦、利亚姆·麦克奈特、安德雷·尼库林（主编）、阿莱克塞·佩古舍夫、亚历山大·皮佩尔斯基、王伊琳（埃莉西娅·沃纳）、米莱娜·韦内娃、鲍里斯·伊奥姆丁。

简体中文文本： 刘羽扬。

祝你好运！