

第二十一屆國際語言學奧林匹亞

巴西·巴西利亞·2024年7月23—31日

團體賽題目

詞彙統計學是一套根據詞彙相似度估計各種語言之間的親緣關係遠近的方法。這種方法通常透過專家由其已手動標記完成的單字表，並找出單字表中的哪些單字為同源詞。雖然如此，語言學家有時也將這種方法應用於透過自動化的程序標註生成的單字表。眾多程序中的其中一種以由蘇聯裔以色列籍語言學家阿哈龍·多爾戈波爾斯基在1964年引入的子音分類概念為基礎。

P. p b ɸ β f v	K. k g x ɣ q ɕ χ ɰ	Y. j ɕ (在字根的開頭)	M. m ɱ
T. t d d̥ θ ð t̪	R. r r̥ ɾ l ʎ ʒ ʎ ʎ	W. w ɱ (在字根的開頭)	N. n ɲ ɳ ɳ
S. s z ʃ ʒ ʒ ɕ ɕ j			Q. ʈ ɖ ɟ
H. 所有 h ɣ ɳ ʒ ʎ h ɳ、母音、以及不在字根開頭的 j ɕ w ɱ			

多爾戈波爾斯基的子音分類

以下將展示已標註完成的單字表若干，每個單字表對應的是一個語系。標註透過下標的數字表示。語言學家以這些單字表為基礎，使用名為「*StarlingNJ*」演算法的兩種簡化版建構語言系譜樹，並為每個單詞分配一個穩定指數。上方的樹狀圖和穩定指數以手動標註的單字表為基礎，而下方的則是基於自動標註的單字表。每種單字表皆透過演算法 A 和演算法 B，分別產出一張樹狀圖，無論其為手動或自動產生。需要注意的是在某些情況下，一個單字表可能可以對應多於一個樹狀圖。在這種情況下，本題將隨機顯示他們其中一個。樹狀圖的每個節點都標有一個數值，代表詞彙統計距離。此距離若越大則代表語言間的親緣關係越近。所以，「反詞彙統計距離」這個術語其實能夠更精準地表達「詞彙統計距離」一詞所欲表達的含義。為便於理解，本題依然採用「詞彙統計距離」一術語。

穩定指數與詞彙統計距離均四捨五入至小數點後兩位。若小數點後第三位小於5，則向下取整，否則向上取整。例如：2.836四捨五入為2.84；0.705四捨五入為0.71；0.703四捨五入為0.70。只有顯示給人類讀者的數值會被四捨五入。換言之，運行演算法的計算機「看見」的是未四捨五入的數值。

請注意，某些單字已知或疑似是外來詞。例如：卡迪維奧語單字 **jok:i**「鹽」來自瓜拉尼語 **juki**；伊派語（梅薩格蘭德方言）單字 **?a:ni**「年」來自西班牙語 **ano**。

在某些情況下，單字表中會列出多個意思相同的詞，並將其以逗號分隔。例如貝羅斯語中的「腳」。

在以下語料中，所有前綴均以「=」符號分隔，所有後綴均以「-」符號分隔。有些單字在使用時需加前綴。這些單字以「=」符號開頭。

所有語料均採用國際音標轉寫。' = 主重音、₁ = 次重音（比主重音弱）、ː = 長音、ɔ̃ = 超短音、XY = X和Y以同一個音發音、ó = 高平調、ò = 低平調、ô = 降調、ʔ = 前聲門化音（發音前喉部短暫阻塞氣流）、' = 外擠音（發音時喉部短暫的阻塞氣流）、◌̥ = 清音、◌̃ = 鼻化音（透過鼻腔發音）、◌̤ = 啞喉聲（低沉、類似沙啞的發音）、ⁿ◌表示鼻腔在發出子音之前有氣流通過、◌^h = 送氣音（發音時有一道氣流噴出）、◌^w = 唇化子音（發音時圓唇）、◌^j = 硬顎化音（發音時舌頭的一部分接近硬顎）。a、æ、ɛ、ɪ、i、ɔ̃、u、ə、ʌ、ɒ、ə、y、ø、ø 為母音。其他特殊字符均為輔音。

△ 與題目中所涉及的語言的相關知識不會為解題帶來任何優勢。

第壹部分. 瓜伊庫魯語系（阿根廷、巴西、巴拉圭）

	托巴語（東部方言）	皮拉加語	莫科維語（查科方言）	卡迪維奧語
雲	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
火	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
魚	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-dʒegi ₃
頭	=qajk ₁	=ʔajk ₁	=qaik ₁	=ak:ilo ₂
殺	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
月亮	ʔawoʒok ₁	ʔa'woʒok ₁	ʃirajɣo ₂	ep:enaj ₃
鼻子	=mik ₁	=ʔmik ₁	=mik ₁	=m:iq:o ₁
鹽	towe ₁	ol'ɣek ₂	ʔwe ₁	jok:i ₁
石頭	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
舌頭	=atʃ-aʒat ₁	=a'ʃ-aʒat ₁	=oʔley-aʒan-aʒat ₂	=ok:el:i ₃

	A 演算法	B 演算法	
手動	<p>詞彙統計距離</p>		<p>穩定性指數：</p> <p>雲 0.50 火 0.50 魚 0.50 頭 0.75 殺 1.00 月亮 0.50 鼻子 1.00 鹽 0.67 石頭 0.75 舌頭 0.50</p>
自動			<p>穩定性指數：</p> <p>雲 0.50 火 0.50 魚 0.75 頭 0.75 殺 1.00 月亮 0.50 鼻子 1.00 鹽 0.25 石頭 0.75 舌頭 0.50</p>

第貳部分. 努比亞語系（埃及、蘇丹）

	敦戈拉維語	克努茲語	迪靈語	卡達魯語	德布里語	比爾吉德語
殺	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
月亮	u'n-at-t₁	an-at-ti₁	nən-ti₁	nən-tu₁	nən-to₁	ma:l₂
水	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
給	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
好	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
風	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
頭髮	'dɪl-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
肚子	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
睡覺	'nɛ:r₁	ne:r₁	ɟer₁	dwallɛli₂	ɟer-i₁	ne:r-i₁
太陽	'masɪl₁	masɪl₁	ɛɟ₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	A 演算法	B 演算法	
手動			殺 0.50 月亮 0.83 水 1.00 給 1.00 好 0.50 風 0.50 頭髮 0.83 肚子 0.83 睡覺 0.83 太陽 0.50
自動			殺 0.33 月亮 0.50 水 0.50 給 0.67 好 0.50 風 0.50 頭髮 0.83 肚子 1.00 睡覺 0.50 太陽 0.50

- (A) (2分) 子音 **ɟ** 音似法語的 *r*，發音部位在舌頭後端。該子音屬於多爾戈波爾斯基分類中的哪一種？你是如何得出這個結論的？
- (B) (2分) 左上角的努比亞語系樹狀圖是該演算法和該標注方式下，兩種可能的樹狀圖之一。畫出另一種可能的樹狀圖。
- (C) (2分) 左下角的努比亞語系樹狀圖是該演算法和該標注方式下，兩種可能的樹狀圖之一。畫出另一種可能的樹狀圖。
- (D) (2分) 如同本題所呈現的其他的數值，右上角樹狀圖根節點的距離0.49只四捨五入至小數點後兩位。四捨五入前的確切數值應為何？

第參部分.馬塔瓜亞語系（阿根廷、玻利維亞、巴拉圭）

	維芝語（貝爾梅霍河下游方言）	維芝語（里瓦達維亞方言）	貝羅斯語	文納耶克語	伊約瓦雅語	曼惠語	尼瓦克勒語（下游方言）	尼瓦克勒語（上游方言）	馬卡語
火	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʰwat ₂	ʔeite ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
魚	ʔwihat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
腳	=patʃu ₁	=qol ₂	=patʃo ₁ , =kala ₂	=pa:kʔo ₁	=ʔsat ₃	=kaʔla ₂	=φo ₄	=φo ₄	=fʔi ₅
水	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweli ₃
給	=ʔwen _{-u} ₁	=wen _{-u} ₁	=ʔwen _{-o} ₁	=ʔwen _{-o} ₁	=ʔwehn-aʔm ₂	=ʔhaj ₃ , =ʔwen ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
好	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
風	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔhlahwu ₄	ʔhlahwu ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
樹	haʔlo ₁	hal ₁	haʔla ₁	haʔla ₁	ʔaʔla ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔkla ₁	naxka-k ₃
頭髮	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔateʔtj ₃	=jeʔs ₄	=ʔewkux-its ₅
殺	=lon ₁	=lon ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=kla ₁	=kla ₁	=lan ₁

	A 演算法	B 演算法	
手動			<p>穩定性指數:</p> <ul style="list-style-type: none"> 火魚 0.78 魚腳 1.00 水給 0.33 好風 0.78 樹髮 0.44 殺 0.89 0.33 0.78 0.67 1.00
自動			<p>穩定性指數:</p> <ul style="list-style-type: none"> 火魚 0.78 魚腳 0.44 水給 0.33 好風 0.56 樹髮 0.67 殺 0.89 0.22 0.67 0.67 1.00

第肆部分. 蒙古語系（中國、蒙古、俄國）

(E) (10分) 請仔細研究以下單字表。請計算出與自動標註和手動標記相對應的穩定指數。

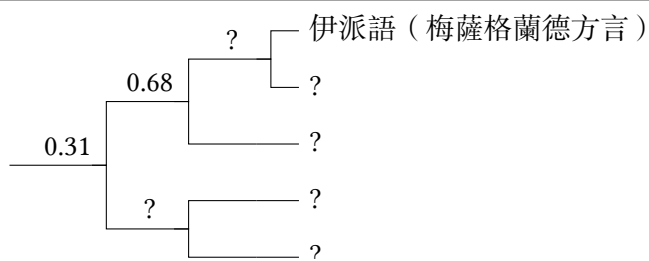
為協助作答，我們將單字「所有」的兩個穩定指數皆提供於此。它們分別為：0.36、0.40（以亂序排列）。

	達斡爾語 (海拉爾 方言)	哈米尼干 語(滿洲 方言)	布里亞特 語(豁裡 方言)	新巴爾虎 語	額魯特語	和碩特語	卡爾梅克 語	喀爾喀語	鄂爾多斯 語	東部裕固 語	保安語
所有	hɔ:₁	bɔlt₂	bɔxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamăġ-₁	pyyyte₄, xamukᵃ-₁	tʃᵃuq₅	hanə-₂
樹皮	hails₁	qalihon₁	χoltɔhɔn₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xɔɣtᵃɔs₂	turusu₃	χalsən₁	arasun₄
肚子	ke:li₁	gɔtəhɔn₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwɣij-₁	ketysy₂	ketesən₂	kele₁
鳥	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃɔr₂	bendzer₂
火	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
路	terg-u:l₁	qargɔi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-lɔə₃	tsam-₁	tʃam-₁	mør₄	mor₄
鹽	hata:₁	dawhɔn₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
游泳	unpa-du₁	ɔmba-₁	tᵃamar-₂	umb-₁	sele-₃	umba-₁	us-tɛi-₄, ø:m-₅	siɣi-₃	usu-tʃᵃi-la-₄	umpa-₁	mba-₁
水	ɔsɔ₁	ɔxɔn₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ɔsɔ₁	usun₁	qᵃusun₁	sə₁
風	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkji₂	salʃkn₂	saɣxi₂	khᵃi:₁	khᵃi:₁	ki₁

第五部分. 尤瑪語系（墨西哥、美國）

(F) (8分) 請仔細研究以下單字表。以下可以看到的是一個基於相同的單字表所構建的樹狀圖。其中有部分數據不提供（語言名稱和詞彙統計距離）。請填寫空缺。請說明該樹狀圖是透過手動或是自動標記產生的，以及其中涉及 A 演算法還是 B 演算法。

	莫哈韋語	科科帕語	亞瓦派語	蒂派語（哈穆爾方言）	伊派語（梅薩格蘭德方言）
短	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuɲ ₁	mə=put-k ₃
鳥	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ='ʃa ₂	ʔa:=ʃa:₂
骨頭	ɲ=a=s=ak ₁	'ɲ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
乾燥	i=ro:-v-k ₁	's=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
肉	kʷi:kʷay ₁	ʔi='ma:tʃ ₂	'kʷe:=ʔo-β-a ₃	'kʷak ₄	kukʷa:j-p ₁
脖子	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:=ʔuk ₂	i:=ʔuk ₂
看到	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
尾巴	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə='juʎ ₂
二	havik-k ₁	'x=wak ₁	'hʷák-i ₁	xə='wak ₁	xə=wak ₁
年	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ^j -₁



(G) (20分) 在此又另外為尤瑪語系生成了一些樹狀圖：這些樹狀圖根部節點的詞彙統計距離（即樹狀圖中最左邊的詞彙統計距離）如下：

1. 0.20
2. 0.23
3. 0.24

請將這些樹狀圖畫出。請針對每個樹狀圖說明其為透過手動或是自動標記產生的，以及其中涉及 A 演算法還是 B 演算法。

(H) (3分) 小題 (G) 裡的詞彙統計距離之中，有兩個是在四捨五入至小數點後兩位得到的：其中，0.23 是從 0.225 四捨五入後得到的。另一個在四捨五入後得到的距離是哪一個？此距離在四捨五入前的具體數值是多少？

(I) (4分) 說明穩定性指數是如何計算出的。

(J) (5分) 說明詞彙統計距離是如何計算出的。

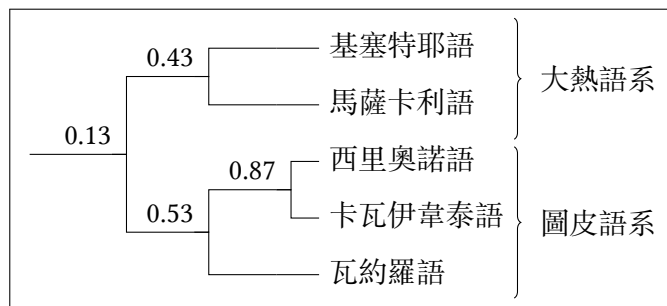
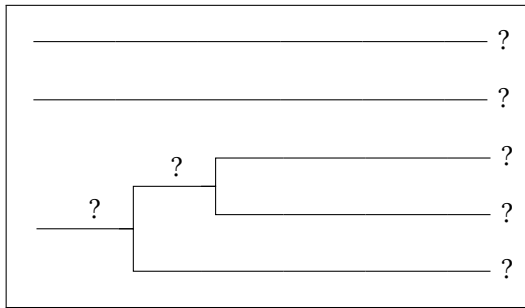
(K) (4分) 說明 A 演算法及 B 演算法之間的區別。

第陸部分. 大熱語系、圖皮語系 (巴西、玻利維亞)

(L) (28分) 大熱語系和圖皮語系是南美洲的兩個主要語系。部分語言學家認為它們存在遙遠的親緣關係。請仔細研究以下單字表。

	A	B	Γ	Δ	E
樹皮	e='e-ke	h ^w i='k ^h Λ	kup='pε	mīβm='tεaj	= 'pε
肚子	'e=rje	=t ^h igi	=ã'ün	= 'tæj	=rε'wek
血	e='ruki	=ka' ⁿ bɾo	=d̄z=a'u	= 'hεβp	=ru'i
燒	= 'rai	=rɔ='k ^h ɹ̄s̄	=po'k ^w a	mū=...='haβp	=ra'pi
脂肪	e='kira	=t ^h wəmi	= 'd̄z=ap	= 'tuβp	= 'kap
腳	'e=i	= 'h ^w aji	= 'βi	=pv'ta	= 'pi
手	'e=o	=ɲi'k ^h ɹ̄a	= 'βo	= 'ɲiβm	= 'pɔ
重	e='usi	=wi't ^h i	=po'ti	=βp'təj	=pɔ'ij
肝臟	'e=ja	= 'nba	=pi'a	=tεiβpkā'nāj	=pi'ʔa
新	e='jasu	= 'ndiwi	=pa'gop	= 'tiβp	=pia'u
根	e='rao	=ja'ɾe	kup=kujɔ'pε	mīβm=ɲiβm=tεa'tiə	=ra'pɔ
皮膚	'e=i	= 'k ^h Λ	= 'pε	= 'tεaj	= 'pit
尾巴	e='rokoĩ	= 'nbi	=d̄z=o'k ^w aj	=nā:='kiβp	= 'raj
白色	'e=fi	=ja'k ^h a	=d̄zi'ra	=βp'dou	= 'sɪŋ
翅膀	e='heo	=ja'ɾa	=pε'o	=ɲi'māu	=pε'pɔ, =ji'wa

以下可以看到的是兩個基於相同的單字表所構建的樹狀圖。其中有部分數據不提供 (語言名稱和詞彙統計距離)。請填寫空缺。請針對每個樹狀圖說明其為透過手動或是自動標記產生的, 以及其中涉及 A 演算法還是 B 演算法。



A	B	Γ	Δ	E
?	?	?	?	?

⚠ 本小題不提供手動標註和穩定指數。

(M) (10分) 以多爾戈波爾斯基類為基礎的自動化程序可能會產生錯誤的結果。這個例子展示的是一個自動化程序的結果，並顯示西里奧諾語與大熱語系中的其中一個語言（基塞特耶語）的距離反而西里奧諾語與圖皮語系其他語言的距離高。請創造一種新版本的自動化程序，並對其簡短地說明。該程序應有辦法正確處理和分類給定的大熱語系和圖皮語系語料。

⚠ 本小題只有在前幾名的隊伍同分時才列入計分。

作者感謝 Alejandra Vidal、Maria Konoshenko、Ilya Gruntov、Jamthô Suyá 回答有關特定語言的問題。
——*Andrey Nikulin、Milena Veneva*

編者：戴誼凡 (Ivan Derzhanski)（技術編輯）、竇脩 (Hugh Dobbs)、Stanislav Gurevich、Boris Iomdin、Liam McKnight、Andrey Nikulin（主編）、Aleksejs Peguševs、Jan Petr、Alexander Piperski、Maria Rubinstein、Milena Veneva、王伊琳 (Elysia Warner)。

正體中文文本：潘同樂。

加油！