

Dwudziesta pierwsza Międzynarodowa Olimpiada Lingwistyczna

Brasília (Brazylia), 23–31 lipca 2024

Zadanie turnieju drużynowego

Leksykostatystyka to zbiór metod mających za zadanie umożliwienie oszacowania stopnia pokrewieństwa języków w oparciu o ich słownictwo. W typowym przypadku metody te stosuje się do długich list wyrazów ręcznie anotowanych przez ekspertów, którzy wskazują pary słów uważanych za wywodzące się ze wspólnego źródła. Czasami jednak językoznawcy stosują metody leksykostatystyczne do list anotowanych przy pomocy automatyzowanych procedur. Jedną z procedur oparta jest na koncepcji *klas spółgłoskowych*, zaproponowanej przez radziecko-izraelskiego językoznawcę Arona Dołgopolskiego w 1964 roku.

P.	p b β φ β f v	K.	k g x γ q ɠ χ u	Y.	j ç (na początku rdzenia)	M.	m ɱ
T.	t d dʰ θ ð t̪ d̪	R.	r r̥ ʀ ɹ l l̥ ʎ ʎ̥	W.	w ɱ (na początku rdzenia)	N.	n ɲ ɳ ɳ̥
S.	s z ʃ ʒ ʂ ʐ ʑ ʑ̥					Q.	ṯ ḏ ḏ̥
H.	ħ ʕ ɦ ʕ̥ ʔ h ɦ ʔ, samogłoski oraz j ç w ɱ (z wyłączeniem początku rdzenia)						

Klasy spółgłoskowe Dołgopolskiego

Poniżej znajdziecie anotowane fragmenty list wyrazów z kilku rodzin językowych z całego świata. Anotacje podano za pomocą cyfr w indeksie dolnym. W oparciu o podane listy za pomocą tzw. algorytmu *StarlingNj* skonstruowano drzewa genealogiczne rodzin językowych. Każdemu wyrazowi przypisano *indeks stabilności*. Drzewa i indeksy stabilności u góry bazują na listach anotowanych ręcznie, zaś dolne – na listach z automatyzowaną anotacją. Przy tym każdej liście odpowiadają dwa drzewa, oparte odpowiednio na algorytmach A i B. Zwróćcie uwagę, że w niektórych przypadkach jednej liście wyrazów może odpowiadać kilka możliwych drzew; w takich przypadkach wybrano losowo jedną z istniejących możliwości. Każdemu węzłowi na każdym drzewie przyporządkowana jest odległość leksykostatystyczna. Im wyższa jest jej wartość, tym bliżej spokrewnione są języki. Trafniej byłoby zatem mówić nie o „odległości leksykostatystycznej”, a o „odwrotnej odległości leksykostatystycznej”. Dla uproszczenia posługujemy się w tym zadaniu pojęciem „odległość leksykostatystyczna”.

Podobnie jak indeksy stabilności, odległości leksykostatystyczne zaokrąglone zostały do dwóch miejsc po przecinku. Jeżeli trzecia cyfra po przecinku jest mniejsza niż 5, zaokrąglamy w dół, w przeciwnym razie – w górę. Przykładowo, 2,836 zaokrąglą się do 2,84, 0,705 do 0,71, a 0,703 do 0,70. Zaokrąglanie stosuje się wyłącznie do wartości przedstawianych odczytującemu je człowiekowi. Oznacza to, że komputer wykonujący algorytm „widzi” wartości niezaokrąglone.

Zwróćcie uwagę, że niektóre wyrazy stanowią znane lub prawdopodobne zapożyczenia z innych języków. Przykład: wyraz **jok:i** ‘sól’ w języku kadiwéu jest zapożyczeniem guarańskiego **juki**, a **?a:ni** ‘rok’ w ipai (Mesa Grande) pochodzi od hiszpańskiego **‘año**.

W niektórych przypadkach do jednego znaczenia przypisano kilka synonimów, oddzielonych przecinkiem. Przykładem może tu być ‘stopa’ w języku vejoz.

W poniższym zbiorze danych wszystkie przedrostki oddzielono znakiem „=”, a przyrostki – znakiem „-”. Niektóre wyrazy występują zawsze z przedrostkami. Poprzedzone są one znakiem „=”.

Dane zapisano w transkrypcji za pomocą Międzynarodowego Alfabetu Fonetycznego. ^ˈ = akcent główny, _ˌ = akcent poboczny (słabszy niż akcent główny), ː = długa głoska, ˚ = bardzo krótka głoska, X̣Y = X i Y wymawiane są razem, tj. jako pojedyncza głoska, ˊ = ton wysoki, ˋ = ton niski, ˆ = ton opadający, ˚ˊ = głoska preglotalizowana (poprzedzona krótką blokadą przepływu powietrza w krtani), ˊ˚ =

spółgłoska eiektywna (wymawiana poprzez gwałtowne zwanie krtani), ɔ̃ = spółgłoska bezdźwięczna, ɔ̃ = głoska nosowa (wymawiana przez nos), ɔ̃ = wymowa skrzypiąca (niski, skrzeczący głos), ɔ̃^o oznacza przepływ powietrza przez nos poprzedzający spółgłoskę, ɔ̃^h = spółgłoska przydechowa (wymawiana z gwałtownym wydechem powietrza), ɔ̃^w = spółgłoska labializowana (tj. wymawiana z zaokrągleniem warg), ɔ̃^j = spółgłoska palatalizowana (tj. miękka). **ɑ, æ, ɛ, ɪ, i, ɔ, ʊ, u, ə, ʌ, ɒ, ɘ, ɤ, ɸ** są samogłoskami. Pozostałe znaki specjalne oznaczają spółgłoski.

△ Znajomość żadnego z języków pojawiających się w zadaniu nie daje przewagi przy jego rozwiązywaniu.

Część I. Rodzina guaicuru (Argentyna, Brazylia, Paragwaj)

	toba (wschodni)	pilagá	mocoví (Chaco)	kadiwéu
chmura	l=ʔok ₁	ʔlo=ʔok ₁	naweyelek ₂	lol:adi ₃
ogień	nodek ₁	ʔd=oleʔ ₂	norek ₁	n=ol:edi ₂
ryba	njaq ₁	ʔnijaq ₁	naʔin ₂	nij:ogo-ɖʒegi ₃
głowa	=qajk ₁	ʔqajk ₁	=qaik ₁	=ak:ilo ₂
zabić	=alawat ₁	=aʔla:t ₁	=alawat ₁	=el:owadi ₁
księżyc	ʔawoʔojk ₁	ʔaʔwoʔojk ₁	ʔirajyo ₂	ep:enaj ₃
nos	=mik ₁	ʔmik ₁	=mik ₁	=m:iq:o ₁
sól	towe ₁	olʔyek ₂	ʔwe ₁	jok:i ₁
kamień	qaʔ ₁	ʔqaʔ ₁	qaʔ ₁	wet:iga ₂
język	=aʔj-aʔat ₁	=aʔj-aʔat ₁	=oʔley-aʔan-aʔat ₂	=ok:el:i ₃

	algorytm A	algorytm B	
<p>ręczne</p> <p>↑ odległość leksykostatystyczna</p>			<p>Indeksy stabilności:</p> <p>chmura 0,50 ogień 0,50 ryba 0,50 głowa 0,75 zabić 1,00 księżyc 0,50 nos 1,00 sól 0,67 kamień 0,75 język 0,50</p>
<p>automatyzowane</p>			<p>Indeksy stabilności:</p> <p>chmura 0,50 ogień 0,50 ryba 0,75 głowa 0,75 zabić 1,00 księżyc 0,50 nos 1,00 sól 0,25 kamień 0,75 język 0,50</p>

Część II. Rodzina nubijska (Egipt, Sudan)

	dongolański	kenuski	dilling	kadaru	debri	birgid
zabić	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
księżyc	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
woda	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
dać	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
dobry	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
wiatr	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
włosy	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
brzuch	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
spać	'nɛ:r₁	ne:r₁	jer₁	dwalleli₂	jer-i₁	ne:r-i₁
słońce	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	algorytm A	algorytm B	
ręczne			Indeksy stabilności: zabić 0,50 księżyc 0,83 woda 1,00 dać 1,00 dobry 0,50 wiatr 0,50 włosy 0,83 brzuch 0,83 spać 0,83 słońce 0,50
automatyzowane			Indeksy stabilności: zabić 0,33 księżyc 0,50 woda 0,50 dać 0,67 dobry 0,50 wiatr 0,50 włosy 0,83 brzuch 1,00 spać 0,50 słońce 0,50

- (A) (2 punkty) Spółgłoska **ɛ** wymawiana jest jak francuskie *r*, za pomocą tylnej części języka. Do której klasy Dołgopolskiego należy i w jaki sposób to ustaliliście?
- (B) (2 punkty) Nubijskie drzewo w lewym górnym rogu jest jednym z dwóch możliwych dla danej kombinacji algorytmu i typu anotacji. Narysujcie drugie możliwe drzewo.
- (C) (2 punkty) Nubijskie drzewo w lewym dolnym rogu jest jednym z dwóch możliwych dla danej kombinacji algorytmu i typu anotacji. Narysujcie drugie możliwe drzewo.
- (D) (2 punkty) Odległość leksykostatystyczna 0,49, przypisana do korzenia drzewa nubijskiego w prawym górnym rogu, została zaokrąglona do dwóch miejsc po przecinku, podobnie jak niektóre inne odległości w niniejszym zadaniu. Ile wynosi dokładna odległość?

Część III. Rodzina mataguayańska (Argentyna, Boliwia, Paragwaj)

	wichí (dolny bieg Bermejo)	wichí (Rivadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaclé (shichaam lhavos)	nivaclé (chisham-nee lhavos)	maká
ogień	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʔhwat ₂	ʔeite ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
ryba	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
stopa	=patʃu ₁	=qolob ₂	=patʃo ₁ , =kala ₂	=pa:kʔoʔ ₁	=ʔsat ₃	=kaʔlaʔ ₂	=ʔoʔ ₄	=ʔoʔ ₄	=fʔiʔ ₅
woda	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
dać	=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
dobry	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔeis ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
wiatr	ʔinwok ^w ₁	ʔinwok ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔhlahwuʔ ₄	ʔhlahwuuʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
drzewo	haʔlo ₁	halob ₁	haʔla ₁	haʔlaʔ ₁	ʔaʔlaʔ ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
włosy	=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
zabić	=lon ₁	=lon ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=kla ₁	=kla ₁	=lan ₁

	algorytm A	algorytm B	
ręczne	<p> wichí (dolny bieg Bermejo) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaclé (shichaam lhavos) nivaclé (chishamnee lhavos) maká </p>	<p> wichí (dolny bieg Bermejo) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaclé (shichaam lhavos) nivaclé (chishamnee lhavos) maká </p>	Indeksy stabilności: ogień 0,78 ryba 1,00 stopa 0,33 woda 0,78 dać 0,44 dobry 0,89 wiatr 0,33 drzewo 0,78 włosy 0,67 zabić 1,00
automatyzowane	<p> wichí (dolny bieg Bermejo) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaclé (shichaam lhavos) nivaclé (chishamnee lhavos) maká </p>	<p> wichí (dolny bieg Bermejo) wichí (Rivadavia) vejoz 'weenhayek iyojwa'aja' manjui nivaclé (shichaam lhavos) nivaclé (chishamnee lhavos) maká </p>	Indeksy stabilności: ogień 0,78 ryba 0,44 stopa 0,33 woda 0,56 dać 0,67 dobry 0,89 wiatr 0,22 drzewo 0,67 włosy 0,67 zabić 1,00

Część IV. Rodzina mongolska (Chińska Republika Ludowa, Mongolia, Federacja Rosyjska)

(E) (10 punktów) Zapoznajcie się z poniższą listą wyrazów. Obliczcie indeksy stabilności odpowiadające anotacji ręcznej i automatyzowanej.

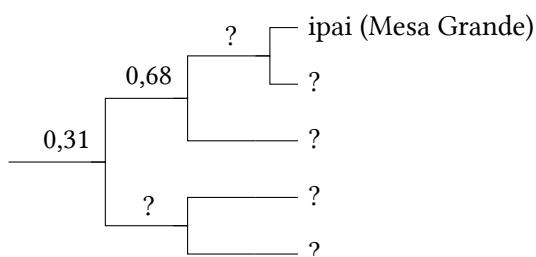
Aby ułatwić Wam zadanie, obliczyliśmy już obydwa indeksy stabilności dla wyrazu ‘wszystkie’. W przypadkowym porządku wynoszą one 0,36 oraz 0,40.

	dagurski (Hailar)	chamni- gański (dial. man- dżurski)	buriacki (dial. cho- ryński)	nowo- bargucki	olocki	choszucki	kałmucki	chałchaski	ordoski	szera- jögurski	baoański
wszystkie	hə:1	bölt2	buxi:3	bygd4	tsug5	lug5	tsuk5, xamak-1	pux3, pugt4, xamäg-1	pyyyte4, xamuk ^h -1	t ^h uq5	hanə-2
kora	hāils1	qalihön1	χoltöhön2	xal ^h u:1	xolts2	xalis1	dursn3	xəłt ^h šs2	turusu3	χalsən1	arasun4
brzuch	ke:li1	getəhən2	gedehen2	gedy:2	ge:s2	gets2	gesn2	gitšs2, xiwłij-1	ketysy2	ketesən2	kele1
ptak	dəgi-1	eiwan1	ʃubu:n1	ʃuwu:1	ʃuvu:1	ʃuwu:1	ʃowun1	ʃuwu1	ʃuβu:1	ʃu:n1, peltʃər2	bendžer2
ogień	gali1	gal1	gal1	gal1	gal1	gal1	gal1	gal1	qal1	qal1	χal1
droga	terg-u:l1	qargöi2	χargi2, zam-1	zam-1	dzam-1	dzam-1	xa:-læ3	tsam-1	tjam-1	mør4	mor4
sól	hata:1	dawhön2	dabhan2	dawuhu:2	daws2	daws2	dawsn2	tawsă2	taβusu2	ta:psən2	dabsuŋ2
pływać	unpa-du1	umba-1	t ^h amar-2	umb-1	sele-3	umba-1	us-tēi-4, ø:m-5	siłi-3	usu-t ^h i-la-4	umpa-1	mba-1
woda	ɔsə1	uxön1	uhan1	u:ha1	usn1	us1	usn1	üsö1	usun1	q ^h usun1	sə1
wiatr	kein1	halkin2	halxin2	halxi2	sal ^h xin2	salk ⁱ 2	sal ^h kn2	salžx ⁱ 2	k ^h i:1	k ^h i:1	ki1

Część V. Rodzina jumańska (Meksyk, Stany Zjednoczone)

(F) (8 punktów) Zapoznajcie się z poniższą listą wyrazów. Poniżej znajduje się drzewo stworzone w oparciu o tę samą listę. Część danych (nazwy języków i odległości leksykostatystyczne) została pominięta. Uzupełnijcie luki. Określcie, czy drzewo sporządzono metodą ręczną czy automatyzowaną oraz czy wygenerowano je przy wykorzystaniu algorytmu A czy B.

	mohave	kokopa	yavapai	tipai (Jamul)	ipai (Mesa Grande)
krótki	wena=wen-a ₁	'xɬ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
ptak	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:2
kość	n=a=s=ak ₁	'n=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
suchy	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
mięso	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
szyja	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:=puk ₂
widzieć	i=ju:-k ₁	'wi:2	'ʔu:1	'wi:w ₂	ə=wu:w ₂
ogon	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
dwa	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
rok	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=tʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 punktów) Dla języków jumańskich wygenerowano kilka dalszych drzew, z następującymi odległościami leksykostatystycznymi w korzeniach drzew (tj. na lewym końcu każdego drzewa):

- 0,20
- 0,23
- 0,24

Narysujcie każde z tych drzew. Dla każdego z drzew określcie, czy zostało sporządzone metodą ręczną czy automatyzowaną oraz czy wygenerowano je przy wykorzystaniu algorytmu A czy B.

(H) (3 punkty) Dwie odległości wymienione w poleceniu (G) zostały zaokrąglone do dwóch miejsc po przecinku: 0,23 uzyskano poprzez zaokrąglenie 0,225. Jaka jeszcze odległość zaokrąglono i ile wynosi jej dokładna wartość?

(I) (4 punkty) Opiszcie, jak obliczane są indeksy stabilności.

(J) (5 punktów) Opiszcie, jak obliczane są odległości leksykostatystyczne.

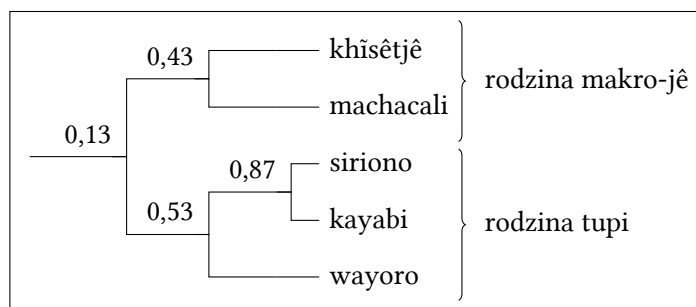
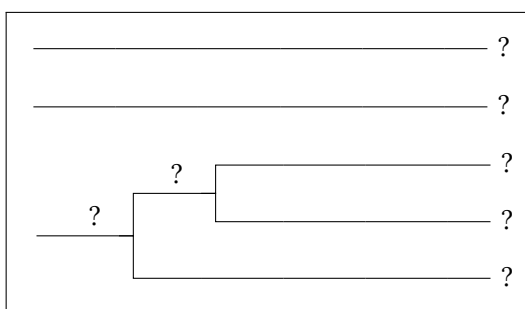
(K) (4 punkty) Wy tłumaczcie różnicę między algorytmem A i algorytmem B.

Część VI. Rodzina makro-jê i rodzina tupi (Brazylia, Boliwia)

(L) (28 punktów) Języki makro-jê i języki tupi stanowią dwie duże rodziny językowe w Ameryce Południowej. Niektórzy językoznawcy uważają je za odległe spokrewnione. Zapoznajcie się z poniższymi listami wyrazów.

	A	B	Γ	Δ	E
kora	e='e-ke	h ^w i='k ^h Λ	kup='pe	mĩβ̃m='tɛaj	= 'pe
brzuch	'e=rje	= 't ^h igi	=ã'ün	= 'tæj	=rɛ'wɛk
krw	e='ruki	=ka'nbɔ	=d̃z=a'ɥ	= 'hɛβ̃p	=ru'ĩ
palić	= 'raĩ	=rɔ='k ^h ɔ̃	=po'k ^w a	mũ=...='haβ̃p	=ra'pi
tuszcz	e='kira	= 't ^h wəmi	= 'd̃z=ap	= 'tuβ̃p	= 'kap
stopa	'e=i	= 'h ^w aji	= 'β̃i	=po'ta	= 'pi
ręka	'e=o	=nĩ'k ^h ɔ̃	= 'β̃o	= 'nĩβ̃m	= 'pɔ
ciężki	e='usi	=wi't ^h ĩ	=po'ti	=β̃p'təj	=pɔ'ij
wątroba	'e=ja	= 'nba	=pi'a	=tɛiβ̃pkĩ'nāj	=pi'ʔa
nowy	e='jasu	= 'ndiwi	=pa'gop	= 'tiβ̃p	=pi'a'u
korzeń	e='rao	=ja'ɽe	kup=kujɔ'pe	mĩβ̃m=nĩβ̃m=tɛa'tiə	=ra'pɔ
skóra	'e=i	= 'k ^h Λ	= 'pe	= 'tɛaj	= 'pit
ogon	e='rokoi	= 'nbi	=d̃z=o'k ^w aj	=nã:='kiβ̃p	= 'raj
biały	'e=ʃĩ	=ja'k ^h a	=d̃zi'ra	=β̃p'douɥ	= 'sĩŋ
skrzydło	e='heo	=ja'ɽa	=pe'o	=nĩ'māuɥ	=rɛ'pɔ, =ji'wa

Poniżej znajdują się dwa drzewa stworzone w oparciu o te same listy. Część danych (nazwy języków i odległości leksykostatystyczne) została pominięta. Uzupełnijcie luki. Dla każdego z drzew określcie, czy zostało sporządzone metodą ręczną czy automatyzowaną oraz czy wygenerowano je przy wykorzystaniu algorytmu A czy B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ Ręczne anotacje i indeksy stabilności zostały w tym poleceniu celowo pominięte.

(M) (10 punktów) Automatyzowane procedury oparte na klasach Dołgopolskiego mogą dawać nieprawidłowe wyniki. W tym przykładzie automatyzowana procedura wykrywa więcej podobieństw pomiędzy siriono i pewnym językiem makro-jê (khîsêtjê) niż pomiędzy siriono a innymi językami tupi. Zaproponujcie i *krótco* opiszcie taką wersję automatyzowanej procedury, która dałaby prawidłową klasyfikację przy zastosowaniu do list wyrazów z rodzin makro-jê i tupi.

⚠ Odpowiedź na to polecenie będzie sprawdzona tylko w przypadku remisu drużyn z najwyższym wynikiem.

Autorzy dziękują Alejandrze Vidal, Marii Konoszenko, Ilji Gruntowowi i Jamthô Suyá za odpowiedzi na pytania dotyczące poszczególnych języków. —Andriej Nikulin, Milena Wenewa

Redakcja: Iwan Derzanski (red. techn.), Hugh Dobbs, Stanisław Gurewicz, Boris Iomdin, Eimear McKnight, Andriej Nikulin (red. odp.), Aleksejs Peguševs, Jan Petr, Aleksandr Piperski, Maria Rubinsztejn, Elysia Warner, Milena Wenewa.

Tekst polski: Przemysław Podleśny.

Powodzenia!