

# Двадесет и първа Международна олимпиада по лингвистика

Бразилия (Бразилия), 23–31 юли 2024 г.

## Задача за отборното състезание

Лексикостатистиката представлява група от методи, чрез които се оценява доколко близко родствени са едни или други езици въз основа на речниковия им състав. Тези методи обикновено се прилагат към дълги списъци с думи, анотирани на ръка от експерти, които посочват дали за някоя двойка думи се смята, че произхожда от един и същ източник. Понякога обаче лингвистите прилагат лексикостатистически методи към списъци с думи, анотирани с помощта на автоматизирани процедури. Една такава процедура се основава на идеята за *класове съгласни*, въведена от съветско-израелския лингвист Аарон Долгополски през 1964 г.

P.	p b β f v	K.	k g x y q ɕ x ɟ	Y.	j ɟ (в началото на корена)	M.	m ɱ
T.	t d θ ð ʈ ɖ	R.	r ɾ ɹ l ʎ ʝ ʎ ʎ	W.	w ɱ (в началото на корена)	N.	n ɲ ɳ ɳ
S.	s z ʒ ʒ ʒ ɕ ɕ ɕ					Q.	ʈ ɖ
H.	ħ ʕ ɳ ʕ ʕ ɳ ʕ ʕ, гласни и j ɟ w ɱ (освен в началото на корена)						

## Класове съгласни на Долгополски

По-долу ще намерите анотирани части от списъци с думи на няколко езикови семейства от света. Анотациите са дадени с долни индекси. Въз основа на тези списъци с помощта на две опростени версии на така наречения алгоритъм *StarlingNj* са конструирани родословни дървета на езиците и на всяка дума е присвоен *индекс на стабилност*. Дърветата и индексите на стабилност в горната част се основават на експертно анотирани списъци с думи, а тези в долната част се основават на списъци, които са анотирани автоматизирано. За всеки списък с думи има конструирани две дървета — по едно с всяка от двете версии на алгоритъма: алгоритъм А и алгоритъм Б. Обърнете внимание, че в някои случаи на един списък с думи съответстват няколко възможни дървета; в такива случаи произволно е избрано само едно дърво. На всеки възел на всяко дърво има приписано лексикостатистическо разстояние. Колкото по-голямо е разстоянието, толкова по-близко родствени са езиците. Поради това би било по-точно да говорим не за „лексикостатистическо разстояние“, а за „обратно лексикостатистическо разстояние“. За простота в тази задача използваме термина „лексикостатистическо разстояние“.

Както индексите на стабилност, така и лексикостатистическите разстояния са закръглени до втория знак след десетичната запетая. Ако третата цифра след десетичната запетая е по-малка от 5, се закръгля надолу; в противен случай се закръгля нагоре. Например 2,836 се закръгля до 2,84, 0,705 — до 0,71, а 0,703 — до 0,70. Закръгляването се прилага само за стойностите, които се показват на хората. С други думи, компютърът, който изпълнява алгоритмите, „вижда“ незакръглените стойности.

Обърнете внимание, че за някои думи е известно или се предполага, че са заети от други езици. Например *jok:i* ‘сол’ на езика кадивео произлиза от думата на гуарани *juki*, а *ʔa:n* ‘година’ на ипай (Меса Гранде) — от испанското *año*.

В някои случаи за едно значение са дадени няколко синонима, разделени със запетая. Един такъв пример е ‘крак’ на езика вехос.

В данните по-долу всички представки са отделени със знак „=“, а всички наставки са отделени със знак „-“. Някои думи не се използват без представки. Те започват със знак „=“.

Данните са записани с помощта на международната фонетична азбука. ' = главно ударение, ˊ = вторично ударение (по-слабо от главното), ː = дълъг звук, ˘ = много кратък звук, XY = X и Y се произнасят като един звук, ˆ = висок тон, ˜ = нисък тон, ˆ = низходящ тон, ˚ = прегло-тализирана съгласна (предшествана от кратко прекъсване на въздушния поток в гърлото), ˚ = абруптивна съгласна (произнася се с кратко прекъсване на въздушния поток в гърлото), ˚ = беззвучна съгласна, ˚ = носов звук (произнася се през носа), ˚ = ларингализация (нисък, скър-цащ звук), ˚ показва, че преди съгласната през носа преминава въздух, ˚<sup>h</sup> = придихателна съгласна (произнася се с придихание), ˚<sup>w</sup> = лабиализирана съгласна (произнася се със закръг-лени устни), ˚<sup>j</sup> = омекотена съгласна. **ɑ, æ, ε, ɪ, i, ə, ʊ, ʉ, ə, ʌ, ɒ, ɔ, ʊ, ɐ, ø** са гласни звукове. Останалите специални символи означават съгласни.

⚠ Познаването на кой да е от посочените в задачата езици не дава предимство при ре-шаването на задачата.

### Част I. Семейство гуайкуру (Аржентина, Бразилия, Парагвай)

	тоба (източен)	пилага	мокови (Чако)	кадивео
облак	l=ʔok <sub>1</sub>	'lo=ʔok <sub>1</sub>	naweyelek <sub>2</sub>	lol:adi <sub>3</sub>
огън	nodek <sub>1</sub>	'd=oleʔ <sub>2</sub>	norek <sub>1</sub>	n=ol:edi <sub>2</sub>
риба	njaq <sub>1</sub>	'nijaq <sub>1</sub>	naʎin <sub>2</sub>	nij:ogo-ɗʒegi <sub>3</sub>
глава	=qajk <sub>1</sub>	'qajk <sub>1</sub>	=qaik <sub>1</sub>	=ak:ilo <sub>2</sub>
убивам	=alawat <sub>1</sub>	=a'la:t <sub>1</sub>	=alawat <sub>1</sub>	=el:owadi <sub>1</sub>
луна	ʔawoʋojk <sub>1</sub>	ʔa'woʋojk <sub>1</sub>	ʃirajʋo <sub>2</sub>	ep:enaj <sub>3</sub>
нос	=mik <sub>1</sub>	'mik <sub>1</sub>	=mik <sub>1</sub>	=m:iq:o <sub>1</sub>
сол	towe <sub>1</sub>	ol'yek <sub>2</sub>	ʔwe <sub>1</sub>	jok:i <sub>1</sub>
камък	qaʔ <sub>1</sub>	'qaʔ <sub>1</sub>	qaʔ <sub>1</sub>	wet:iga <sub>2</sub>
език	=atʃ̣-awat <sub>1</sub>	=a'tʃ̣-aʃat <sub>1</sub>	=oʔley-awan-awat <sub>2</sub>	=ok:el:i <sub>3</sub>

	алгоритъм А	алгоритъм Б	
експертно	<p>↑ лексикостатистическо разстояние</p>		Индекси на стабилност: облак 0,50 огън 0,50 риба 0,50 глава 0,75 убивам 1,00 луна 0,50 нос 1,00 сол 0,67 камък 0,75 език 0,50
автоматизирано			Индекси на стабилност: облак 0,50 огън 0,50 риба 0,75 глава 0,75 убивам 1,00 луна 0,50 нос 1,00 сол 0,25 камък 0,75 език 0,50

### Част II. Нубийско семейство (Египет, Судан)

	донголави	кенузи	дилинг	кадару	дебри	биргид
убивам	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
луна	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
вода	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
давам	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
добър	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
вятър	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
коса (косми)	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
корем	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
спя	'nɛ:r₁	ne:r₁	jer₁	dwallɛli₂	jer-i₁	ne:r-i₁
слънце	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	алгоритъм А	алгоритъм Б	
експертно			Индекси на стабилност:
			убивам 0,50 луна 0,83 вода 1,00 давам 1,00 добър 0,50 вятър 0,50 коса (косми) 0,83 корем 0,83 спя 0,83 слънце 0,50
автоматизирано			Индекси на стабилност:
			убивам 0,33 луна 0,50 вода 0,50 давам 0,67 добър 0,50 вятър 0,50 коса (косми) 0,83 корем 1,00 спя 0,50 слънце 0,50

- (A) (2 точки) Съгласната **ѡ** се произнася като грасирано («френско») *p*, т. е. в най-задната част на езика. Към кой клас на Долгополски принадлежи и как го установихте?
- (B) (2 точки) Нубийското дърво в горния ляв ъгъл е само едно от две възможни дървета за тази комбинация на алгоритъм и тип аотиране. Нарисувайте другото възможно дърво.
- (C) (2 точки) Нубийското дърво в долния ляв ъгъл е само едно от две възможни дървета за тази комбинация на алгоритъм и тип аотиране. Нарисувайте другото възможно дърво.
- (D) (2 точки) Лексикостатистическото разстояние 0,49, означено до корена на нубийското дърво, дадено в горния десен ъгъл, както и някои други разстояния в тази задача, е закръглено до два знака след десетичната запетая. Какво е точното разстояние?

**Част III. Матагвайско семейство (Аржентина, Боливия, Парагвай)**

	уичи (долно- бермехски)	уичи (Ривада- вия)	вехос	уенхайек	ийоухааха	манхуй	нивакле (шичаам хлавос)	нивакле (чишамнее хлавос)	мака
огън	ʔitox <sub>1</sub>	ʔitəx <sub>1</sub>	ʔitah <sub>1</sub>	ʔi:tax <sub>1</sub>	ʔhwat <sub>2</sub>	ʔeite <sub>1</sub>	ʔitax <sub>1</sub>	ʔitax <sub>1</sub>	feʔt <sub>2</sub>
риба	ʔwahat <sub>1</sub>	wahat <sub>1</sub>	wahat <sub>1</sub>	ʔwa:hat <sub>1</sub>	siʔjus <sub>-1</sub>	ʃiʔjus <sub>-1</sub>	saxetʃ <sub>-1</sub>	saxetʃ <sub>-1</sub>	sehets <sub>-1</sub>
крак	=patʃu <sub>1</sub>	=qəɓ <sub>2</sub>	=patʃo <sub>1</sub> , =kala <sub>2</sub>	=pa:kʔoʔ <sub>1</sub>	=ʔsat <sub>3</sub>	=kaʔlaʔ <sub>2</sub>	=foʔ <sub>4</sub>	=foʔ <sub>4</sub>	=fʔiʔ <sub>5</sub>
вода	ʔinot <sub>1</sub>	ʔinət <sub>1</sub>	wah <sub>2</sub>	ʔina:t <sub>1</sub>	ʔiʔnat <sub>1</sub>	ʔaʔnat <sub>1</sub>	jinaʔt <sub>1</sub>	jinaʔt <sub>1</sub>	iweliʔ <sub>3</sub>
давам	=ʔweŋ-u <sub>1</sub>	=weŋ-u <sub>1</sub>	=ʔweŋ-o <sub>1</sub>	=ʔweŋ-oʔ <sub>1</sub>	=ʔwehn-aʔm <sub>2</sub>	=ʔhajʔ <sub>3</sub> , =ʔweŋ <sub>2</sub>	=xut <sub>4</sub>	=xut-ej <sub>4</sub>	tis-ix <sub>5</sub>
добър	ʔis <sub>1</sub>	ʔis <sub>1</sub>	ʔis <sub>1</sub>	ʔis <sub>1</sub>	ʔes <sub>1</sub>	ʔeis <sub>1</sub>	ʔis <sub>1</sub>	ʔis <sub>1</sub>	t=ejkʔun-ej <sub>2</sub>
вятър	ʔinwok <sup>w</sup> <sub>1</sub>	ʔinwək <sub>1</sub>	ʔihwok <sup>w</sup> <sub>1</sub>	=ja:ʔ <sub>2</sub> , =x <sup>w</sup> ox <sup>w</sup> <sub>3</sub>	ʔhlahwuʔ <sub>4</sub>	ʔhlahwu <sup>w</sup> <sub>4</sub>	ʔaβiʔm <sub>5</sub>	ʔaβiʔm <sub>5</sub>	tʔunikʔi <sub>6</sub>
дърво	haʔlo <sub>1</sub>	halə <sub>1</sub>	haʔla <sub>1</sub>	haʔlaʔ <sub>1</sub>	ʔaʔlaʔ <sub>1</sub>	ʔaʔla-k <sub>1</sub>	ʔaʔkxi-juk <sub>2</sub>	jiʔklaʔ <sub>1</sub>	naxka-k <sub>3</sub>
коса (кос- ми)	=ʔwule-j <sub>1</sub>	=wule-j <sub>1</sub>	=ʔwole-j <sub>1</sub>	=ʔwo:le-ç <sub>1</sub> , hi:lenaç <sub>2</sub>	=ʔwole <sub>1</sub>	=ʔwole-j <sub>1</sub>	=ʔateʔtʃ <sub>3</sub>	=jeʔs <sub>4</sub>	=ʔewkux-its <sub>5</sub>
убивам	=lon <sub>1</sub>	=lən <sub>1</sub>	=lan <sub>1</sub>	=la:ŋ <sub>1</sub>	=ʔlaʔan <sub>1</sub>	=ʔlan <sub>1</sub>	=kla <sup>w</sup> <sub>1</sub>	=kla <sup>w</sup> <sub>1</sub>	=lan <sub>1</sub>

	алгоритъм А	алгоритъм Б	
експертно			Индекси на стабилност: огън 0,78 риба 1,00 крак 0,33 вода 0,78 давам 0,44 добър 0,89 вятър 0,33 дърво 0,78 коса (косми) 0,67 убивам 1,00
автоматизирано			Индекси на стабилност: огън 0,78 риба 0,44 крак 0,33 вода 0,56 давам 0,67 добър 0,89 вятър 0,22 дърво 0,67 коса (косми) 0,67 убивам 1,00

### Част IV. Монголско семейство (Китайска народна република, Монголия, Русия)

(Е) (10 точки) Разгледайте следния списък с думи. Изчислете индексите на стабилност, съответстващи както на експертните, така и на автоматизираните анотации.

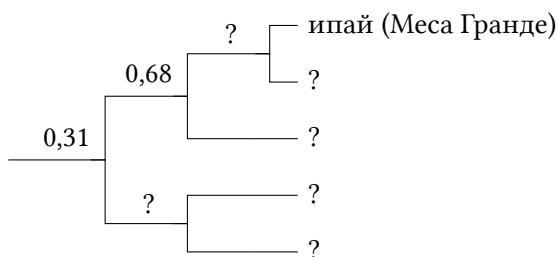
За да облекчим работата Ви, вече сме изчислили двата индекса на стабилност за думата ‘всички’. В случаен ред те са: 0,36 и 0,40.

	даурски (хайларски)	хамнигански (манджурски)	бурятски (хорински)	новобаргутски	уулдски	хошутски	калмикски	халхамонголски	ордоски	шираюгурски	бонански
всички	hə:₁	bolt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak-₁	pux₃, pugt₄, xamāg-₁	pyyute₄, xamuk <sup>h</sup> -₁	tʃ <sup>h</sup> uq₅	hanə-₂
кора	hails₁	qalihon₁	χoltəhən₂	xalʃhu:₁	xolts₂	xalis₁	dursn₃	xəɮt <sup>h</sup> ʒs₂	turusu₃	χalsən₁	arasun₄
корем	ke:li₁	getəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitʃs₂, xiwɮʃij-₁	ketysy₂	ketesən₂	kele₁
птица	dəgi-₁	eiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendʒer₂
огън	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
път	terg-u:l₁	qargūi₂	χargi₂, zam-₁	zam-₁	dzam-₁	dzam-₁	xa:-ləθ₃	tsam-₁	tʃam-₁	mər₄	mor₄
сол	hata:₁	dawhən₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
плувам	unpa-du₁	umba-₁	t <sup>h</sup> amar-₂	umb-₁	sele-₃	umba-₁	us-təi-₄, ø:m-₅	siɮi-₃	usu-tʃ <sup>h</sup> i-la-₄	umpa-₁	mba-₁
вода	əɕə₁	uxən₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	usö₁	usun₁	q <sup>h</sup> usun₁	sə₁
вятър	kein₁	halkin₂	halxin₂	halxi₂	salʃxin₂	salkʃi₂	salʃkn₂	saɮxǰ₂	k <sup>h</sup> i:₁	k <sup>h</sup> i:₁	ki₁

### Част V. Юманско семейство (Мексико, САЩ)

(F) (8 точки) Разгледайте следния списък с думи. По-долу можете да видите дърво, изградено въз основа на същия списък. Някои данни (имена на езици и лексикостатистически разстояния) са изпуснати. Запълнете празните места. Посочете дали дървото е експертно или автоматизирано, както и дали е генерирано с алгоритъм А или Б.

	мохаве	кокопа	явапай	типай (Хамул)	ипай (Меса Гранде)
къс	wena=wen-a <sub>1</sub>	'xɬ=ʔut <sub>2</sub>	'tʃkr=ot-i <sub>2</sub>	lə=ʔuɲ <sub>1</sub>	mə=put-k <sub>3</sub>
птица	ʔitʃ=i=jer <sub>1</sub>	'ʃa <sub>2</sub>	'ʔ=ʔʃ=sa <sub>2</sub>	aʔ=ʃa <sub>2</sub>	ʔa:=ʃa:2
кост	ɲ=a=s=ak <sub>1</sub>	'ɲ=j=a:k <sub>1</sub>	'tʃ=j=a:k-a <sub>1</sub>	'ak <sub>1</sub>	aq <sub>1</sub>
сух	i=ro:-v-k <sub>1</sub>	'ʃ=ʔar <sub>2</sub>	'ru-β-i <sub>1</sub>	's=ʔa:j <sub>3</sub>	sa:j <sub>3</sub>
месо	k <sup>w</sup> i:k <sup>w</sup> ay <sub>1</sub>	ʔi='ma:tʃ <sub>2</sub>	'k <sup>w</sup> e:='θo-β-a <sub>3</sub>	'k <sup>w</sup> ak <sub>4</sub>	kuk <sup>w</sup> a:j-p <sub>1</sub>
шия, врат	maʎaqe <sub>1</sub>	'm=puk <sub>2</sub>	'mlq <sub>1</sub>	i:='puk <sub>2</sub>	i:='puk <sub>2</sub>
виждам	i=ju:-k <sub>1</sub>	'wi:2	'ʔu:1	'wi:w <sub>2</sub>	ə=wu:w <sub>2</sub>
опашка	i:=ʔar <sub>1</sub>	'ʃ=juʎ <sub>2</sub>	'β=hé <sub>3</sub>	ʃə='juʎ <sub>2</sub>	xə=juʎ <sub>2</sub>
две	havik-k <sub>1</sub>	'x=wak <sub>1</sub>	'h <sup>w</sup> âk-i <sub>1</sub>	xə='wak <sub>1</sub>	xə=wak <sub>1</sub>
година	hu:ðe <sub>1</sub>	'mat-'ka:m <sub>2</sub>	'ʔ=ʔʃ <sup>h</sup> ur-a <sub>3</sub>	mat-'wam <sub>2</sub>	ʔa:n <sup>i</sup> <sub>1</sub>



(G) (20 точки) За юманските езици са генерирани още няколко други дървета със следните лексикостатистически разстояния в основата на дървото (лексикостатистическите разстояния отляво на всяко дърво):

1. 0,20
2. 0,23
3. 0,24

Нарисувайте всяко от тези дървета. За всяко от дърветата посочете дали е експертно или автоматизирано, както и дали е генерирано с алгоритъм А или Б.

(H) (3 точки) Две от разстоянията, изброени в подзадача (G), са закръглени до втория знак след десетичната запетая: 0,23 е закръглено от 0,225. Кое друго разстояние е закръглено и каква е незакръглената му стойност?

(I) (4 точки) Обяснете как се изчисляват индексите на стабилност.

(J) (5 точки) Обяснете как се изчисляват лексикостатистическите разстояния.

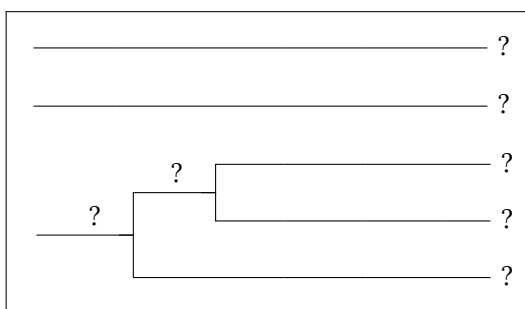
(K) (4 точки) Обяснете разликата между алгоритъм А и алгоритъм Б.

### Част VI. Семейство макроже и тупийско семейство (Бразилия, Боливия)

(L) (28 точки) Макроже и тупийското семейство са две големи езикови семейства в Южна Америка. Някои езиковеди считат, че те са далечно родствени. Разгледайте следните списъци с думи.

	A	B	Г	Δ	E
кора	e='e-ke	h <sup>w</sup> ĩ='k <sup>h</sup> λ	kʉp='pɛ	mĩβ̃m='tɛaj	'pɛ
корем	'e=rje	't <sup>h</sup> igi	=ã'ũn	'tɛj	=rɛ'wɛk
кръв	e='ruki	=ka <sup>n</sup> bɔ	=d̃z=a'ɯ	'hɛβ̃p	=ru'i
горя	'raĩ	=rɔ='k <sup>h</sup> ɔ̃	=po <sup>k</sup> wa	mũ=...='haβ̃p	=ra'pi
мазнина	e='kɪra	't <sup>h</sup> wəmi	=d̃z=ap	'tuβ̃p	'kap
крак	'e=i	'h <sup>w</sup> aji	'βi	=pɔ'ta	'pi
ръка	'e=o	=ɲĩ <sup>k</sup> hɔ	'βo	'ɲĩβ̃m	'pɔ
тежък	e='usi	=wi <sup>t</sup> hĩ	=pɔ'ti	=β̃p'tɛj	=pɔ'ij
черен дроб	'e=ja	'nba	=pi'a	=tɛiβ̃pkĩ'nãj	=pi'ʔa
нов	e='jasu	'ndiwi	=pa'gɔp	'tiβ̃p	=pia'u
корен	e='rao	=jaɾɛ	kʉp=kʉjo'pɛ	mĩβ̃m=ɲĩβ̃m=tɛa'tiã	=ra'pɔ
кожа	'e=i	'k <sup>h</sup> λ	'pɛ	'tɛaj	'pit
опашка	e='rokoĩ	'nbi	=d̃z=o <sup>k</sup> waj	=nã:='kiβ̃p	'raj
бял	'e=ʃĩ	=ja <sup>k</sup> hɔ	=d̃zi'ra	=β̃p'doɯ	'siŋ
крило	e='heo	=ja'ɾa	=pɛ'o	=ɲĩ <sup>m</sup> ãɯ	=pɛ'pɔ, =ji'wa

По-долу можете да видите две дървета, изградени въз основа на същите списъци. Някои данни (имена на езици и лексикостатистически разстояния) са изпуснати. Запълнете празните места. За всяко от дърветата посочете дали е експертно или автоматизирано, както и дали е генерирано с алгоритъм A или B.



A	B	Г	Δ	E
?	?	?	?	?



⚠ Експертните анотации и индексите на стабилност в тази подзадача са умишлено пропуснати.

(М) (10 точки) Автоматизираните процедури, основани на класовете на Долгополски, могат да дават неверни резултати. В разглеждания пример автоматизираната процедура открива повече прилики между сирионо и един от езиците макроже (кхинсетже), отколкото между сирионо и другите тупийски езици. Предложете модифицирана автоматизирана процедура, която би довела до правилна класификация, ако се приложи към горните списъци с думи от езици макроже и тупийски езици, и я опишете *накратко*.

⚠ Тази подзадача ще се оценява само в случай на равенство между отборите с най-високи резултати.

Авторите благодарят на Алехандра Видал, Мария Коношенко, Иля Грунгов и Ямто Суя за консултациите по отделните езици.  
—*Андрей Никулин, Милена Венева*

---

**Редактори:** Милена Венева, Станислав Гуревич, Иван Держански (техн. ред.), Хю Добс, Борис Иомдин, Лиам Макнайт, Андрей Никулин (отг. ред.), Алексей Пегушев, Ян Петър, Александър Пиперски, Мария Рубинщайн, Елисия Уорнър.

**Български текст:** Милена Венева.

Наслука!