

Einundzwanzigste Internationale Linguistik-Olympiade

Brasília (Brasilien), 23.–31. Juli 2024

Aufgabe des Gruppenwettbewerbs

Die Lexiko-statistik ist eine Sammlung von Methoden, die entwickelt wurden, um auf der Grundlage des Wortschatzes abzuschätzen, wie eng verwandt beliebige Sprachen sind. Normalerweise werden diese Methoden auf langen, manuell von Experten annotierten Wortlisten angewendet, wobei die Experten auch angeben, ob zwei bestimmte Wörter vermutlich aus derselben Quelle stammen. Manchmal allerdings verwenden Linguisten lexiko-statistische Methoden mit Wortlisten, die durch automatisierte Verfahren annotiert werden. Ein solches automatisiertes Annotationsverfahren basiert auf dem Konzept der *Konsonantenklassen*, die 1964 vom sowjetisch-israelischen Linguisten Aharon Dolgopolsky eingeführt wurden.

P.	p b ɓ φ β f v	K.	k g x y q ɠ χ u	Y.	j ɟ (am Anfang der Wurzel)	M.	m ŋ
T.	t d d̥ θ ð t̥ d̥	R.	r ɾ ɽ l ʎ ʟ ʎ̥ ʟ̥	W.	w ɱ (am Anfang der Wurzel)	N.	n ɲ ɳ ɲ
S.	s z ʃ ʒ ʂ ʐ z̥ c ɟ					Q.	t̪ d̪
H.	h ʕ ɦ ʔ ʕ h ɦ ʔ, Vokale und j ɟ w ɱ (außer am Anfang der Wurzel)						

Dolgopolskys Konsonantenklassen

Unten sind annotierte Fragmente von Wortlisten aus mehreren Sprachfamilien der Welt. Die Annotationen werden durch tiefgestellte Ziffern dargestellt. Auf der Grundlage dieser Listen wurden Sprachfamilienstambäume mithilfe zweier vereinfachter Versionen des sogenannten *StarlingNJ*-Algorithmus gebaut, und ein *Stabilitätsindex* wurde jedem Wort zugewiesen. Die Bäume und Stabilitätsindizes oben basieren auf manuell generierten Wortlisten, und die unten basieren auf automatisch annotierten Wortlisten. Für jede Wortliste gibt es zwei Bäume, die mithilfe zweier Versionen des Algorithmus – Algorithmus A und Algorithmus B – gebaut wurden. Beachtet, dass in einigen Fällen mehrere mögliche Bäume einer Wortliste entsprechen; in solchen Fällen wurde nur ein Baum zufällig ausgewählt. Jedem Knoten in jedem Baum wird eine lexiko-statistische Distanz zugewiesen. Je höher die Distanz, desto enger die Verwandtschaft zwischen den Sprachen. Deswegen wäre es präziser den Begriff „umgekehrte lexiko-statistische Distanz“ als den Begriff „lexiko-statistische Distanz“ zu verwenden. In dieser Aufgabe wird der Begriff „lexiko-statistische Distanz“ der Einfachheit halber verwendet.

Sowohl die Stabilitätsindizes als auch die lexiko-statistischen Distanzen werden auf zwei Dezimalstellen gerundet. Wenn die dritte Ziffer nach dem Komma kleiner als 5 ist, rundet man ab; andernfalls rundet man auf. Zum Beispiel wird 2,836 auf 2,84, 0,705 auf 0,71, und 0,703 auf 0,70 gerundet. Das Runden gilt nur für die Werte, die menschlichen Lesern gezeigt werden. Anders gesagt kann der Computer, der die Algorithmen ausführt, „sehen“, was die ungerundeten Werte sind.

Beachtet, dass einige Wörter bekanntermaßen oder vermutlich aus anderen Sprachen entlehnt sind. Zum Beispiel wurde das Wort **jok:i** ‚Salz‘ der Kadiwéu-Sprache aus dem Guaraní-Wort **juki** und **?a:n̄i** ‚Jahr‘ des Ipai (Mesa Grande) aus dem spanischen **’ajo** entlehnt.

In einigen Fällen werden in den Wortlisten mehrere Synonyme einer einzigen Bedeutung durch Kommata getrennt angegeben. Ein Beispiel ist ‚Fuß‘ im Vejoz.

In den Daten unten werden alle Präfixe durch das „-“-Zeichen und alle Suffixe durch das „-“-Zeichen getrennt. Einige Wörter kommen nur mit Präfixen vor. Diese Wörter beginnen mit dem Zeichen „-“.

Die Daten werden mit dem Internationalen Phonetischen Alphabet transkribiert. ' = Hauptbetonung, ˌ = Nebenbetonung (schwächer als Hauptbetonung), ː = langer Laut, ˚ = sehr kurzer Laut, ˘Y =

X und Y werden als ein einziger Laut ausgesprochen, \acute{o} = hoher Ton, \grave{o} = tiefer Ton, \hat{o} = fallender Ton, ?o = vorglottalisierter Laut (nach vorhergehender Blockierung des Luftstroms im Hals ausgesprochen), ? = ejektiver Laut (durch eine kurze Blockierung des Luftstroms im Hals ausgesprochen), ? = stimmloser Laut, \tilde{o} = nasalisierter Laut (durch die Nase ausgesprochen), ? = knarrende Stimme (ein tiefer, kratziger Stimmklang), ? zeigt an, dass vor dem Konsonanten etwas Luft durch die Nase strömt, ?^h = aspirierter Konsonant (mit Luftstoß ausgesprochen), ?^w = labialisierter Konsonant (mit gerundeten Lippen ausgesprochen), ?^j = palatisierter Laut (durch Hebung eines Teils der Zunge in Richtung des harten Gaumens ausgesprochen). **a, æ, ε, i, ɨ, ɔ, ʊ, u, ə, ʌ, ɒ, ɘ, y, ø, ø** sind Vokale. Andere Sonderzeichen sind Konsonanten.

⚠ Kenntnisse irgendeiner der genannten Sprachen bringen bei der Lösung dieser Aufgabe keinen Vorteil.

Teil I. Guaikurú-Sprachfamilie (Argentinien, Brasilien, Paraguay)

	Toba (Östlich)	Pilagá	Mokoví (Chaco)	Kadiwéu
Wolke	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
Feuer	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
Fisch	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-dʒegi ₃
Kopf	=qajk ₁	=qajk ₁	=qaik ₁	=ak:ilo ₂
töten	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
Mond	ʔawokojk ₁	ʔa'woʎojk ₁	ʃirajʎo ₂	ep:enaj ₃
Nase	=mik ₁	=mik ₁	=mik ₁	=m:iq:o ₁
Salz	towe ₁	ol'ʎek ₂	ʔwe ₁	jok:i ₁
Stein	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
Zunge	=atʃ-aʎat ₁	=a'tʃ-aʎat ₁	=oʔleʎ-aʎan-aʎat ₂	=ok:el:i ₃

	Algorithmus A	Algorithmus B	
manuell	<p>lexikostatistische Distanz</p>		Stabilitätsindizes: Wolke 0,50 Feuer 0,50 Fisch 0,50 Kopf 0,75 töten 1,00 Mond 0,50 Nase 1,00 Salz 0,67 Stein 0,75 Zunge 0,50
automatisiert			Stabilitätsindizes: Wolke 0,50 Feuer 0,50 Fisch 0,75 Kopf 0,75 töten 1,00 Mond 0,50 Nase 1,00 Salz 0,25 Stein 0,75 Zunge 0,50

Teil II. Nubische Sprachfamilie (Ägypten, Sudan)

	Dongolawi	Kenuzi	Dilling	Kadaru	Debri	Birgid
töten	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
Mond	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
Wasser	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	eji₁
geben	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
gut	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛɲ₂	azze-n₃
Wind	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
Haare	'dil-ti₁	sir₂	tɛl-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
Bauch	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
schlafen	'nɛ:r₁	ne:r₁	ɟɛr₁	dwallɛli₂	ɟɛr-i₁	ne:r-i₁
Sonne	'masil₁	masil₁	ɛɟ₂	aju₂	ɛɟgal-to₃	ʔi:zi₂

	Algorithmus A	Algorithmus B	
manuell			Stabilitätsindizes: töten 0,50 Mond 0,83 Wasser 1,00 geben 1,00 gut 0,50 Wind 0,50 Haare 0,83 Bauch 0,83 schlafen 0,83 Sonne 0,50
automatisiert			Stabilitätsindizes: töten 0,33 Mond 0,50 Wasser 0,50 geben 0,67 gut 0,50 Wind 0,50 Haare 0,83 Bauch 1,00 schlafen 0,50 Sonne 0,50

- (A) (2 Punkte) Der Konsonant **ɟ** wird wie *r* im Wort *rot* ausgesprochen, am hinteren Teil der Zunge. Zu welcher Dolgopolsky-Klasse gehört dieser Konsonant, und wie habt ihr das festgestellt?
- (B) (2 Punkte) Der nubische Baum oben links ist nur einer von zwei möglichen Bäumen für diese Kombination von Algorithmus und Annotationstyp. Zeichnet den anderen möglichen Baum.
- (C) (2 Punkte) Der nubische Baum unten links ist nur einer von zwei möglichen Bäumen für diese Kombination von Algorithmus und Annotationstyp. Zeichnet den anderen möglichen Baum.
- (D) (2 Punkte) Wie andere Distanzen in dieser Aufgabe wurde die (der Wurzel des nubischen Baumes oben rechts zugewiesene) lexikostatistische Distanz 0,49 auf zwei Dezimalstellen gerundet. Was ist die exakte Distanz?

Teil III. Mataguayische Sprachfamilie (Argentinien, Bolivien, Paraguay)

	Wichí (Nieder- Bermejo)	Wichí (Rivada- via)	Vejoz	‘Weenhayek	Iyojwa’aja’	Manjui	Nivaklé (Shichaam Lhavos)	Nivaklé (Chis- hamnee Lhavos)	Maká
Feuer	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	ʔh ^h wat ₂	ʔe ^h it ^h e ₁	ʔitax ₁	ʔitax ₁	feʔt ₂
Fisch	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	siʔjus ₋₁	ʃiʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
Fuß	=patʃ ^h u ₁	=qol ₂	=patʃ ^h o ₁ , =kala ₂	=pa:kʔoʔ ₁	=ʔsat ₃	=kaʔlaʔ ₂	=φoʔ ₄	=φoʔ ₄	=fʔiʔ ₅
Wasser	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔiʔnʔat ₁	ʔaʔnat ₁	jinaʔt ₁	jinaʔt ₁	iweliʔ ₃
geben	=ʔweŋ ^h -u ₁	=weŋ ^h -u ₁	=ʔweŋ ^h -o ₁	=ʔweŋ ^h -oʔ ₁	=ʔweh ^h n-aʔm ₂	=ʔhajʔ ₃ , =ʔweŋ ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
gut	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	ʔes ₁	ʔe ^h is ₁	ʔis ₁	ʔis ₁	t=ejkʔun-ej ₂
Wind	ʔinwok ^w ₁	ʔinwək ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	ʔh ^h lah ^h wuʔ ₄	ʔh ^h lah ^h wuuʔ ₄	ʔaβiʔm ₅	ʔaβiʔm ₅	tʔunikʔi ₆
Baum	haʔlo ₁	hal ₁	haʔla ₁	haʔlaʔ ₁	ʔaʔlaʔ ₁	ʔaʔla-k ₁	ʔaʔkxi-juk ₂	jiʔklaʔ ₁	naxka-k ₃
Haare	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lənax ₂	=ʔwole ₁	=ʔwole-j ₁	=ʔateʔtj ₃	=jeʔs ₄	=ʔewkux-its ₅
töten	=lon ₁	=lən ₁	=lan ₁	=la:ŋ ₁	=ʔlaʔan ₁	=ʔlan ₁	=kla ₁	=kla ₁	=lan ₁

	Algorithmus A	Algorithmus B	
manuell			Stabilitätsindizes: Feuer 0,78 Fisch 1,00 Fuß 0,33 Wasser 0,78 geben 0,44 gut 0,89 Wind 0,33 Baum 0,78 Haare 0,67 töten 1,00
automatisiert			Stabilitätsindizes: Feuer 0,78 Fisch 0,44 Fuß 0,33 Wasser 0,56 geben 0,67 gut 0,89 Wind 0,22 Baum 0,67 Haare 0,67 töten 1,00

Teil IV. Mongolische Sprachfamilie (Volksrepublik China, Mongolei, Russland)

(E) (10 Punkte) Analysiert die folgende Wortliste. Berechnet die Stabilitätsindizes, die den manuellen und den automatisierten Annotationen entsprechen.

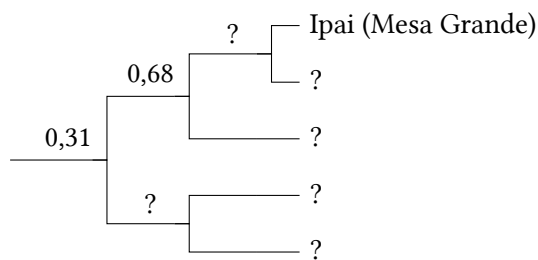
Als Hilfestellung wurden beide Stabilitätsindizes für das Wort ‚alle‘ schon berechnet. In zufälliger Reihenfolge sind diese Indizes 0,36 und 0,40.

	Dagurisch (Hailar)	Chamnigan (Man- dschu)	Burjatisch (Chori)	Neu- Bargutisch	Ööld	Choschut	Kalmückisch	Chalcha	Ordos	Ost-Yugur	Bonan
alle	hɔ:₁	bölt₂	buxi:₃	bygd₄	tsug₅	lug₅	tsuk₅, xamak₋₁	pux₃, pugt₄, xamăġ₋₁	pyyyte₄, xamukʰ₋₁	tʃʰuq₅	hanə₂
Rinde	hails₁	qalihon₁	χoltōhōn₂	xalʰhu:₁	xolts₂	xalis₁	dursn₃	xɔɣtʰōs₂	turusu₃	χalsən₁	arasun₄
Bauch	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitis₂, xiwɣij₋₁	ketysy₂	ketesən₂	kele₁
Vogel	dəgi₋₁	eiwan₁	ʃubun₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃər₂	bendzer₂
Feuer	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
Weg	terg-u:l₁	qargūi₂	χargi₂, zam₋₁	zam₋₁	dzam₋₁	dzam₋₁	xa:-lɔə₃	tsam₋₁	tʃam₋₁	mør₄	mor₄
Salz	hata:₁	dawhōn₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsă₂	taβusu₂	ta:psən₂	dabsuŋ₂
schwimmen	unpa-du₁	umba₋₁	tʰamar₋₂	umb₋₁	sele₋₃	umba₋₁	us-təi₋₄, ø:m₋₅	siɣi₋₃	usu-tʃʰi-la₋₄	umpa₋₁	mba₋₁
Wasser	ɔsɔ₁	oxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ʊsö₁	usun₁	qʰusun₁	sə₁
Wind	kein₁	halkin₂	halxin₂	halxi₂	salxin₂	salkji₂	saljkn₂	saɣxi₂	kʰi:₁	kʰi:₁	ki₁

Teil V. Yuma-Sprachfamilie (Mexiko, USA)

(F) (8 Punkte) Analysiert die folgende Wortliste. Unten ist ein Baum, der mit derselben Liste gebaut wurde. Einige Daten (Namen von Sprachen und lexikostatistische Distanzen) fehlen. Füllt die Lücken aus. Gebt an, ob der Baum manuell oder automatisch generiert ist, sowie ob er von Algorithmus A oder B generiert ist.

	Mohave	Kokopa	Yavapai	Tipai (Jamul)	Ipai (Mesa Grande)
kurz	wena=wen-a ₁	'xʌ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
Vogel	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=ʔʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:2
Knochen	ɲ=a=s=ak ₁	'ɲ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
trocken	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
Fleisch	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
Hals	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:='puk ₂	i:=puk ₂
sehen	i=ju:-k ₁	'wi:2	'ʔu:1	'wi:w ₂	ə=wu:w ₂
Schwanz	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə='juʎ ₂	xə=juʎ ₂
zwei	havik-k ₁	'x=wak ₁	'h ^w âk-i ₁	xə='wak ₁	xə=wak ₁
Jahr	hu:ðe ₁	'mat-'ka:m ₂	'ʔ=ʔʃ ^h ur-a ₃	mat-'wam ₂	ʔa:n ⁱ ₁



(G) (20 Punkte) Einige andere Bäume wurden für die Yuman-Sprachen generiert und zeigen die folgenden lexikostatistischen Distanzen an der Wurzel jedes Baumes (ganz links an jedem Baum):

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

Zeichnet jeden dieser Bäume. Gebt für jeden der Bäume an, ob er manuell oder automatisch generiert ist, sowie ob er von Algorithmus A oder B generiert ist.

(H) (3 Punkte) Zwei Distanzen in Teilaufgabe (G) wurden auf zwei Dezimalstellen gerundet: 0,23 wurde von 0,225 hochgerundet. Welche andere Distanz wurde gerundet, und wie hoch ist ihr präziser Wert?

(I) (4 Punkte) Erklärt, wie die Stabilitätsindizes berechnet werden.

(J) (5 Punkte) Erklärt, wie die lexikostatistische Distanzen berechnet werden.

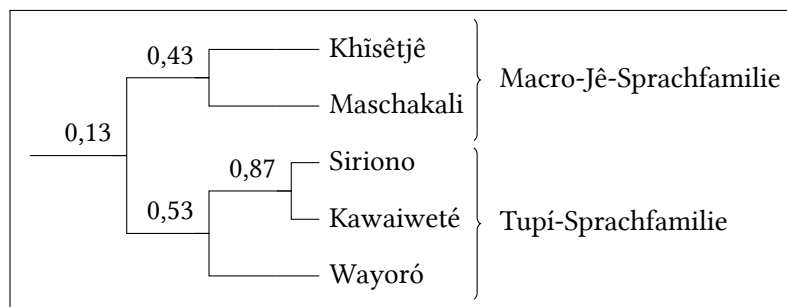
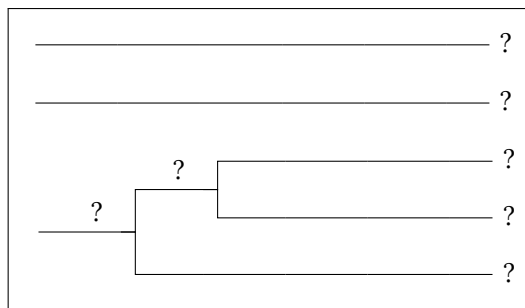
(K) (4 Punkte) Erklärt den Unterschied zwischen den Algorithmen A und B.

Teil VI. Macro-Jê-Sprachfamilie und Tupí-Sprachfamilie (Brasilien, Bolivien)

(L) (28 Punkte) Die Macro-Jê-Sprachen und Tupí-Sprachen sind zwei bedeutende Sprachfamilien Südamerikas. Einige Linguisten denken, dass sie entfernt verwandt sind. Analysiert die folgenden Wortlisten.

	A	B	Γ	Δ	E
Rinde	e='e-ke	h ^w i='k ^h Λ	kup='pε	mīβm='tεaj	= 'pε
Bauch	'e=rje	= 't ^h igi	=ã ün	= 'tæj	=rε'wek
Blut	e='ruki	=ka'nbrɔ	=d̄z=a'u	= 'hεβp	=ru'i
verbrennen	= 'rai	=rɔ='k ^h λš	=po'k ^w a	mũ=...='haβp	=ra'pi
Fett	e='kira	= 't ^h wəmi	= 'd̄z=ap	= 'tuβp	= 'kap
Fuß	'e=i	= 'h ^w aji	= 'βi	=pɔ'ta	= 'pi
Hand	'e=o	=ɲi'k ^h λa	= 'βo	= 'ɲiβm	= 'pɔ
schwer	e='usi	=wi't ^h i	=po'ti	=βp'təj	=pɔ'ij
Leber	'e=ja	= 'nba	=pi'a	=tεiβpk'i'nāj	=pi'ʔa
neu	e='jasu	= 'ndiwi	=pa'gop	= 'tiβp	=pia'u
Wurzel	e='rao	=ja'rε	kup=kujɔ'pε	mīβm=ɲiβm=tεa'tiə	=ra'pɔ
Haut	'e=i	= 'k ^h Λ	= 'pε	= 'tεaj	= 'pit
Schwanz	e='rokoï	= 'nbi	=d̄z=o'k ^w aj	=nā:='kiβp	= 'raj
weiß	'e=ʃi	=ja'k ^h a	=d̄zi'ra	=βp'douɥ	= 'sɪŋ
Flügel	e='heo	=ja'rα	=pε'o	=ɲi'māuɥ	=pε'pɔ, =ji'wa

Unten sind zwei Bäume, die mit denselben Listen gebaut wurden. Einige Daten (Namen von Sprachen und lexikostatistische Distanzen) fehlen. Füllt die Lücken aus. Gebt für jeden der Bäume an, ob er manuell oder automatisch generiert ist, sowie ob er von Algorithmus A oder B generiert ist.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ In dieser Teilaufgabe wurden die manuellen Annotierungen und Stabilitätsindizes absichtlich ausgelassen.

(M) (10 Punkte) Automatisierte Verfahren, die auf Dolgopolsky-Klassen basieren, können ungenaue Ergebnisse erzielen. In diesem Beispiel identifiziert das automatische Verfahren mehr Ähnlichkeiten zwischen dem Siriono und einer bestimmten Macro-Jê-Sprache (Khîsêtjê) als zwischen dem Siriono und anderen Tupí-Sprachen. Schlagt ein geändertes automatisiertes Verfahren vor, das eine richtige Klassifikation erzielen würde, wenn es auf die Macro-Jê- und Tupí-Wortlisten oben angewendet würde, und beschreibt es *kurz*.

⚠ Diese Teilaufgabe wird nur bei Punktgleichheit zwischen den besten Teams bewertet werden.

Die Autoren danken Alejandra Vidal, Maria Konoshenko, Ilya Gruntov und Jamthô Suyá für die Beantwortung ihrer Fragen zu einzelnen Sprachen.
—*Andrey Nikulin, Milena Veneva*

Redaktion: Ivan Derzhanski (tech. Red.), Hugh Dobbs, Stanislav Gurevich, Boris Iomdin, Liam McKnight, Andrey Nikulin (Chefredakteur), Aleksejs Peguševs, Jan Petr, Alexander Piperski, Maria Rubinstein, Milena Veneva, Elysia Warner.

Deutsche Fassung: Elysia Warner.

Viel Erfolg!