

Dvacátá první Mezinárodní olympiáda v lingvistice

Brasília (Brazílie), 23.–31. července 2024

Úloha soutěže družstev

Lexikostatistika je skupina metod navrhnutá pro odhadování toho, jak blízce jsou si jakékoliv jazyky navzájem příbuzné na základě jejich slovní zásoby. Tyto metody jsou obvykle použity na dlouhé seznamy slov ručně okomentovaných experty, kteří označí, jestli konkrétní pár slov domnělaně pochází ze stejného zdroje. Občas však lingvisté použijí lexikostatistické metody na seznamy slov okomentované pomocí automatizovaných procedur. Jedna taková procedura je založená na konceptu *souhláskových tříd*, zavedených sovětsko-izraelským lingvistou Aharonem Dolgopolským v roce 1964.

P.	p b ɓ φ β f v	K.	k g x γ q ɠ χ ɰ	Y.	j ɟ (na počátku kořene)	M.	m ɱ
T.	t d d̥ θ ð t̥ d̥	R.	r r̥ ɽ ɻ l ɭ ʎ ʟ ʠ	W.	w ɱ (na počátku kořene)	N.	n ɲ ɳ ɰ
S.	s z ʃ ʒ ʂ ʐ ʑ ʒ ɟ					Q.	ᶥ ᶧ
H.	ħ ʕ ɦ ʕ ʔ h ɦ ʔ, samohláska a j ɟ w ɱ (s výjimkou počátku kořene)						

Dolgopolského souhláskové třídy

Níže naleznete okomentované fragmenty seznamů slov z několika světových jazykových rodin. Komentáře jsou dány jako číslice v dolním indexu. Na základě těchto seznamů byly sestrojeny rodinné stromy jazyků pomocí dvou zjednodušených verzí takzvaného *StarlingNf* algoritmu, a ke každému slovu byl přiřazen *koeficient stability*. Stromy a koeficienty na vrchu jsou založeny na ručně okomentovaných seznamech slov, a ty na spodu jsou založeny na seznamech, které byly okomentovány automaticky. Pro každý seznam slov jsou sestrojeny dva stromy: pro algoritmus A a algoritmus B. Povšimněte si, že v některých případech je možné sestrojít více různých stromů; v takových případech byl náhodně zvolen jen jeden z nich. Ke každému uzlu na každém stromu je přiřazena lexikostatistická vzdálenost. Čím větší vzdálenost, tím blíže je příbuzenství mezi jazyky. Přesnější než „lexikostatistická vzdálenost“ by tudíž byl pojem „převrácená lexikostatistická vzdálenost“. Pro jednoduchost budeme v této úloze používat pojem „lexikostatistická vzdálenost“.

Jak koeficienty stability, tak lexikostatistické vzdálenosti jsou zaokrouhleny na dvě desetinná místa. Pokud je třetí číslice za desetinnou čárkou menší než 5, zaokrouhlujeme dolů; v opačném případě zaokrouhlujeme nahoru. Například, 2,836 je zaokrouhleno na 2,84, 0,705 je zaokrouhleno na 0,71, a 0,703 je zaokrouhleno na 0,70. Zaokrouhlení se týká pouze hodnot ukázaných lidským čtenářům. Jinak řečeno: počítač, který vykonává algoritmus, „vidí“ nezaokrouhlené hodnoty.

Povšimněte si, že o některých slovech se ví, anebo je předpokládáno, že byla přejmuta z jiných jazyků. Například, slovo **jok:i** ‚sůl‘ jazyka kadiwéu je přejímka od slova **juki** v jazyce guaraní, a **?a:nj** ‚rok‘ jazyka ’iipay (Mesa Grande) je přejmuto ze španělského **'ajno**.

V některých případech je k jednomu významu uvedeno několik synonym oddělených čárkou. Příkladem je ‚chodidlo‘ v jazyce vejoz.

V datech níže jsou předpony odděleny znakem „=“, a všechny přípony jsou odděleny znakem „-“. Některá slova se používají výhradně s předponami. Ta začínají znakem „=“.

Data jsou přepsána pomocí mezinárodní fonetické abecedy. ¹ = hlavní přízvuk, ₁ = vedlejší přízvuk (slabší než hlavní přízvuk), ◌ː = dlouhá hláska, ◌̆ = velmi krátká hláska, X̂Y = X a Y jsou vysloveny jako jedna hláska, ◌ˊ = vysoký tón, ◌ˋ = nízký tón, ◌ˎ = klesavý tón, ◌ˑ = předglotalizovaná hláska (předcházená krátkým zahrazením proudu vzduchu v hrdle), ◌ʰ = ejektivní souhláska (vyslovovaná krátkým zahrazením proudu vzduchu v hrdle), ◌̥ = neznělá hláska, ◌̃ = nasalizovaná hláska (vyslovovaná skrze nos),

Úloha soutěže družstev

◌̥ = třepená fonace (hluboký, skřípavý zvuk), ◌̥ⁿ označuje proudění vzduchu skrze nos před souhláskou, ◌^h = souhláska s přídechem (vyslovená s výdechem vzduchu), ◌^w = labializovaná souhláska (vyslovená se zaokrouhlenými rty), ◌^j = palatalizovaná hláska (vyslovována s jazykem posunutým blíže tvrdému patru). **a, æ, ε, i, ĩ, ɔ, u, ɥ, ə, ʌ, ɒ, ɘ, y, ɐ, ø** jsou samohlásky. Zbylé speciální znaky jsou souhlásky.

△ Znalost jakéhokoliv z jazyků zmíněných v této úloze nepřináší žádnou výhodu při jejím řešení.

Část I. Rodina gvajkurú (Argentina, Brazílie, Paraguay)

	toba (východní)	pilagá	mocoví (Chaco)	kadiwéu
mrak	l=ʔok ₁	'lo=ʔok ₁	naweyelek ₂	lol:adi ₃
oheň	nodek ₁	'd=oleʔ ₂	norek ₁	n=ol:edi ₂
ryba	njaq ₁	'nijaq ₁	naʎin ₂	nij:ogo-dʒegi ₃
hlava	=qajk ₁	'qajk ₁	=qaik ₁	=ak:ilo ₂
zabít	=alawat ₁	=a'la:t ₁	=alawat ₁	=el:owadi ₁
měsíc	ʔawoʔojk ₁	ʔa'woʔojk ₁	ʃirajyo ₂	ep:enaj ₃
nos	=mik ₁	'mik ₁	=mik ₁	=m:iq:o ₁
sůl	towe ₁	ol'yek ₂	ʔwe ₁	jok:i ₁
kámen	qaʔ ₁	'qaʔ ₁	qaʔ ₁	wet:iga ₂
jazyk	=atʃ-akat ₁	=a'tʃ-aʃat ₁	=oʔley-akʌn-akat ₂	=ok:el:i ₃

	algoritmus A	algoritmus B	
ručně	<p>lexikostatistická vzdálenost</p>		Koefficienty stability: mrak 0,50 oheň 0,50 ryba 0,50 hlava 0,75 zabít 1,00 měsíc 0,50 nos 1,00 sůl 0,67 kámen 0,75 jazyk 0,50
automatizované			Koefficienty stability: mrak 0,50 oheň 0,50 ryba 0,75 hlava 0,75 zabít 1,00 měsíc 0,50 nos 1,00 sůl 0,25 kámen 0,75 jazyk 0,50

Část II. Núbijská rodina (Egypt, Súdán)

	dongolawi	kenuzi	dilling	kadaru	debri	birgid
zabít	'bɛ:₁	be:₁	hur₂	wur-i₂	wur-i₂	fila:l-e₁
měsíc	u'n-at-t₁	an-at-ti₁	nɔn-ti₁	nɔn-tu₁	nɔn-to₁	ma:l₂
voda	'ɛss₁	essi₁	ɔti₁	ɔto₁	ɔtu₁	ɛji₁
dát	'tir₁	tir₁	ti₁	ti₁	ti₁	te:-n₁
dobry	'sɛrɛ:₁	sere:₁	ken₂	kɛn₂	kɛŋ₂	azze-n₃
vitr	'turug₁	turug₁	irf-i₂	irf-o₂	irf-o₂	kurr-i₃
vlasý	'dil-ti₁	si:r₂	tel-ti₁	til-tu₁	til-tu₁	ur=dill-e₁
břicho	'tu:₁	tu:₁	te-te₂	to₁	to₁	tu:₁
spát	'nɛ:r₁	ne:r₁	jer₁	dwalleli₂	jer-i₁	ne:r-i₁
slunce	'masil₁	masil₁	ɛj₂	aju₂	ɛŋgal-to₃	ʔi:zi₂

	algoritmus A	algoritmus B	
ručně			Koefficienty stability: zabít 0,50 měsíc 0,83 voda 1,00 dát 1,00 dobrý 0,50 vítr 0,50 vlasý 0,83 břicho 0,83 spát 0,83 slunce 0,50
automatizovaně			Koefficienty stability: zabít 0,33 měsíc 0,50 voda 0,50 dát 0,67 dobrý 0,50 vítr 0,50 vlasý 0,83 břicho 1,00 spát 0,50 slunce 0,50

- (A) (2 body) Souhláska **ɣ** je vyslovována jako francouzské *r*, na zadní části jazyka. Do které Dolgopolského třídy náleží, a jak jste na to přišli?
- (B) (2 body) Núbijský strom vlevo nahoře je pouze jedním ze dvou možných stromů pro tuto kombinaci algoritmu a typu komentáře. Nakreslete druhý možný strom.
- (C) (2 body) Núbijský strom vlevo dole je pouze jedním ze dvou možných stromů pro tuto kombinaci algoritmu a typu komentáře. Nakreslete druhý možný strom.
- (D) (2 body) Lexikostatistická vzdálenost 0,49 (přiřazená kořenu núbijskému stromu vpravo nahoře) byla zaokrouhlena na dvě desetinná místa, stejně jako ostatní vzdálenosti v této úloze. Jaká je přesná vzdálenost?

Část III. Matagvajská rodina (Argentina, Bolívie, Paraguay)

	wichí (dolní Bermejeño)	wichí (Rivadavia)	vejoz	'weenhayek	iyojwa'aja'	manjui	nivaclé (shichaam lhavos)	nivaclé (chisham-nee lhavos)	maká
ohěň	ʔitox ₁	ʔitox ₁	ʔitah ₁	ʔi:tax ₁	'hwat ₂	'ʔeiti _e ₁	ʔitax ₁	ʔitax ₁	fe't ₂
ryba	ʔwahat ₁	wahat ₁	wahat ₁	ʔwa:hat ₁	si'ʔjus ₋₁	ʃi'ʔjus ₋₁	saxetʃ ₋₁	saxetʃ ₋₁	sehets ₋₁
chodidlo	=patʃ _u ₁	=qol _o ₂	=patʃ _o ₁ , =kala ₂	=pa:k'oʔ ₁	=sat ₃	=ka'laʔ ₂	=φoʔ ₄	=φoʔ ₄	=f'iʔ ₅
voda	ʔinot ₁	ʔinot ₁	wah ₂	ʔina:t ₁	ʔi'n'iat ₁	ʔa'ʔnat ₁	jina't ₁	jina't ₁	iweliʔ ₃
dát	=ʔwen _u ₋₁	=wen _u ₋₁	=ʔwen _o ₋₁	=ʔwen _o ₋₁	=wɛhn-aʔm ₂	=hajʔ ₃ , =wɛn ₂	=xut ₄	=xut-ej ₄	tis-ix ₅
dobrý	ʔis ₁	ʔis ₁	ʔis ₁	ʔis ₁	'ʔes ₁	'ʔeis ₁	ʔis ₁	ʔis ₁	t=ejk'un-ej ₂
vítr	ʔinwok ^w ₁	ʔinwok ₁	ʔihwok ^w ₁	=ja:ʔ ₂ , =x ^w ox ^w ₃	'hlahwuʔ ₄	'hlahwu ^w ʔ ₄	ʔaβi'm ₅	ʔaβi'm ₅	t'unik'i ₆
strom	ha'lo ₁	hal _o ₁	ha'la ₁	ha'laʔ ₁	ʔa'laʔ ₁	ʔa'la-k ₁	ʔa'kxi-juk ₂	ji'klaʔ ₁	naxka-k ₃
vlasy	=ʔwule-j ₁	=wule-j ₁	=ʔwole-j ₁	=ʔwo:le-ç ₁ , hi:lenaç ₂	=ʔwole ₁	=ʔwole-j ₁	=ʃate'ʔʃ ₃	=je's ₄	=ʔewkux-its ₅
zabít	=lon ₁	=lon ₁	=lan ₁	=la:ŋ ₁	=laʔan ₁	=lan ₁	=kla _n ₁	=kla _n ₁	=lan ₁

	algoritmus A	algoritmus B	
ručně			Koefficienty stability: oheň 0,78 ryba 1,00 chodidlo 0,33 voda 0,78 dát 0,44 dobrý 0,89 vítr 0,33 strom 0,78 vlasy 0,67 zabít 1,00
automatizovaně			

Část IV. Mongolská rodina (Čínská lidová republika, Mongolsko, Rusko)

(E) (10 bodů) Prozkoumejte následující seznam slov. Spočítejte koeficienty stability odpovídající jak ručním, tak automatizovaným komentářům.

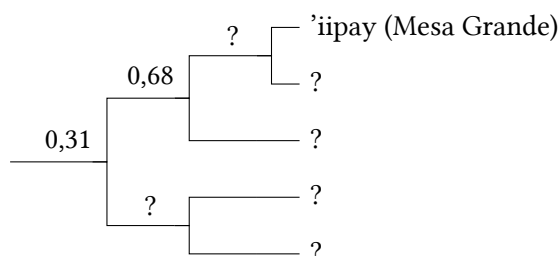
Abychom Vám pomohli, spočítali jsme oba koeficienty stability pro slovo ‚vše‘. V náhodném pořadí se jedná o 0,36 a 0,40.

	dagurština (hailarský dialekt)	chamnigan- ština (man- džuský dialekt)	burjatština (chorijský dialekt)	novo- bargut- ština	öeletština	chošúd- ština	kalmyčtina	chalcha- ština	ordoština	šira-ju- gurština	bonanština
vše	hɔ:₁	bolt₂	boxi:₃	bygd₄	ṭsug₅	lug₅	ṭsuk₅, xamak-₁	pux₃, pugt₄, xamāg-₁	pyyyte₄, xamukᵃ-₁	ṭʰuq₅	hanə-₂
kůra	hāils₁	qalihon₁	χoltəhən₂	xalʰhu:₁	xolts₂	xalis₁	dursn₃	xəɮtᵃšs₂	turusu₃	χalsən₁	arasun₄
břicho	ke:li₁	gətəhən₂	gedehen₂	gedy:₂	ge:s₂	gets₂	gesn₂	gitšs₂, xiwɮij-₁	ketysy₂	ketesən₂	kele₁
pták	dəgi-₁	əiwan₁	ʃubu:n₁	ʃuwu:₁	ʃuvu:₁	ʃuwu:₁	ʃowun₁	ʃuwu₁	ʃuβu:₁	ʃu:n₁, peltʃor₂	bendžer₂
oheň	gali₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	gal₁	qal₁	qal₁	χal₁
cesta	terg-u:l₁	qargōi₂	χargi₂, zam-₁	zam-₁	ḏzam-₁	ḏzam-₁	xa:-lkə₃	ṭsam-₁	ṭjam-₁	mør₄	mor₄
sůl	hata:₁	dawhən₂	dabhan₂	dawuhu:₂	daws₂	daws₂	dawsn₂	tawsā₂	taβusu₂	ta:psən₂	dabsuŋ₂
plavat	unpa-du₁	umba-₁	tᵃamar-₂	umb-₁	sele-₃	umba-₁	us-təi-₄, ø:m-₅	sikʃi-₃	usu-tʰi-la-₄	umpa-₁	mba-₁
voda	əɕə₁	uxon₁	uhan₁	u:ha₁	usn₁	us₁	usn₁	ʊsš₁	usun₁	qᵃusun₁	sə₁
vítr	kein₁	halkin₂	halxin₂	halxi₂	salʰxin₂	salkʰi₂	salʰkn₂	saɮxī₂	kᵃi:₁	kᵃi:₁	ki₁

Část V. Jumanská rodina (Mexiko, Spojené státy americké)

(F) (8 bodů) Prozkoumejte následující seznam slov. Níže je k vidění strom, který byl sestaven na základě stejného seznamu. Některá data (názvy jazyků a lexikostatistické vzdálenosti) chybí. Vyplňte mezery. Stanovte, jestli byl strom sestaven ručně, nebo automaticky, a jestli byl vytvořen pomocí algoritmu A, nebo B.

	mojave	cocopa	yavapai	tiipay (jamul)	'iipay (Mesa Grande)
krátký	wena=wen-a ₁	'xɬ=ʔut ₂	'tʃkr=ot-i ₂	lə=ʔuj ₁	mə=put-k ₃
pták	ʔitʃ=i=jer ₁	'ʃa ₂	'ʔ=tʃ=sa ₂	aʔ=ʃa ₂	ʔa:=ʃa:₂
kost	ɲ=a=s=ak ₁	'ɲ=j=a:k ₁	'tʃ=j=a:k-a ₁	'ak ₁	aq ₁
suchý	i=ro:-v-k ₁	'ʃ=ʔar ₂	'ru-β-i ₁	's=ʔa:j ₃	sa:j ₃
maso	k ^w i:k ^w ay ₁	ʔi='ma:tʃ ₂	'k ^w e:=ʔo-β-a ₃	'k ^w ak ₄	kuk ^w a:j-p ₁
krk	maʎaqe ₁	'm=puk ₂	'mlq ₁	i:=ʔuk ₂	i:=puk ₂
vidět	i=ju:-k ₁	'wi:₂	'ʔu:₁	'wi:w ₂	ə=wu:w ₂
ocas	i:=ʔar ₁	'ʃ=juʎ ₂	'β=hé ₃	ʃə=ʔuʎ ₂	xə=juʎ ₂
dvě	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 ^l ₁



(G) (20 bodů) Několik dalších stromů bylo vytvořeno z jumanských jazyků, s následujícími lexikostatistickými vzdálenostmi u kořene stromu (tj. lexikostatistickými vzdálenostmi na levém kraji stromu):

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

Nakreslete každý z těchto stromů. Pro každý ze stromů stanovte, jestli byl sestaven ručně, nebo automaticky, a jestli byl vytvořen pomocí algoritmu A, nebo B.

(H) (3 body) Dvě ze vzdáleností uvedených v Úkolu (G) byly zaokrouhleny na dvě desetinná místa: 0,23 bylo zaokrouhleno z 0,225. Jaká další vzdálenost byla zaokrouhlena, a jaká je její přesná hodnota?

(I) (4 body) Vysvětlete, jak jsou počítány koeficienty stability.

(J) (5 bodů) Vysvětlete, jak jsou počítány lexikostatistické vzdálenosti.

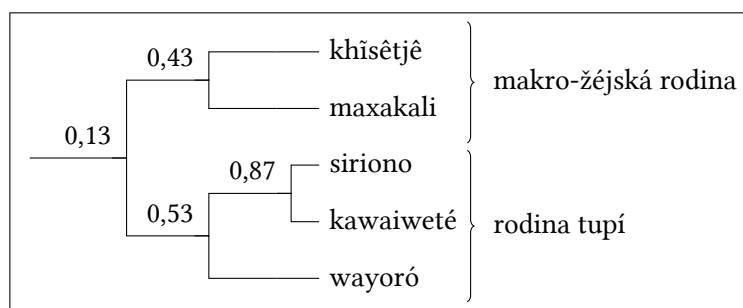
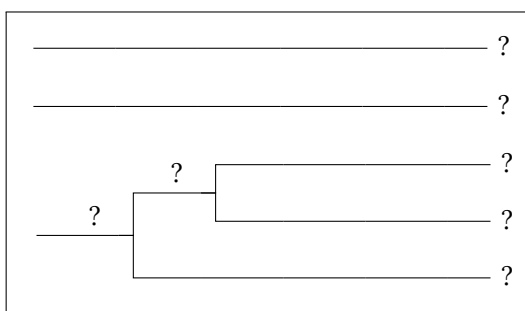
(K) (4 body) Vysvětlete rozdíl mezi algoritmem A a algoritmem B.

Část VI. Makro-žéjská rodina a rodina tupí (Brazílie, Bolívie)

(L) (28 bodů) Makro-žéjská rodina a rodina tupí jsou dvě z hlavních jazykových rodin v Jižní Americe. Někteří lingvisté věří, že jsou vzdáleně spřízněné. Prozkoumejte následující seznamy slov.

	A	B	Γ	Δ	E
kůra	e='e-ke	h ^w i='k ^h Λ	kup='pe	mĩβm='tɛaj	= 'pe
břicho	'e=rje	't ^h igi	=ã'ün	'tæj	=rɛ'wɛk
krev	e='ruki	=ka' ⁿ bɾo	=d̥z=a'u	'hɛβp	=ru'i
pálit	'raĩ	=rɔ='k ^h ɔ̃	=po'k ^w a	mũ=...='haβp	=ra'pi
tuk	e='kira	't ^h wəmi	=d̥z=ap	'tuβp	'kap
chodidlo	'e=i	'h ^w aji	'βi	=pɔ'ta	'pi
ruka	'e=o	=ɲi' ^{k^h} ɔ̃	'βo	'ɲĩβm	'pɔ
těžký	e='usi	=wi't ^h ĩ	=pɔ'ti	=βp'təj	=pɔ'ij
játra	'e=ja	' ⁿ ba	=pi'a	=tɛiβpkĩ'nãj	=pi'ʔa
nový	e='jasu	' ⁿ diwi	=pa'gop	'tiβp	=pia'u
kořen	e='rao	=ja'tɛ	kup=kujɔ'pe	mĩβm=ɲĩβm=tɛa'tiə	=ra'pɔ
kůže	'e=i	'k ^h Λ	'pe	'tɛaj	'pit
ocas	e='rokoi	' ⁿ bi	=d̥z=ɔ'k ^w aj	=nã:'kiβp	'raj
bílý	'e=ʃĩ	=ja' ^{k^h} a	=d̥zi'ra	=βp'douɲ	'sĩɲ
křídlo	e='heo	=ja'tɾa	=pɛ'o	=ɲi'mãuɲ	=pɛ'pɔ, =ji'wa

Níže jsou k vidění dva stromy, které byly sestaveny na základě stejných seznamů. Některá data (názvy jazyků a lexikostatické vzdálenosti) chybí. Vyplňte mezery. Pro každý ze stromů stanovte, jestli byl sestaven ručně, nebo automaticky, a jestli byl vytvořen pomocí algoritmu A, nebo B.



A	B	Γ	Δ	E
?	?	?	?	?

⚠ Ruční komentáře a koeficienty stability byly záměrně vynechány v tomto úkolu.

(M) (10 bodů) Automatizované procedury založené na Dolgopolského třídách mohou vést k nesprávným výsledkům. V tomto příkladě automatizovaná procedura detekuje více podobností mezi jazykem siriono a určitým makro-žéjským jazykem (khîsêtjê), než mezi jazykem siriono a ostatními tupíjskými jazyky. Navrhněte a *stručně* popište upravenou automatizovanou proceduru, která by vedla ke správné klasifikaci, pokud by byla použita na makro-žéjské a tupíjské seznamy slov výše.

⚠ Tento úkol bude ohodnocený pouze v případě remízy mezi týmy s nejvyšším počtem bodů.

Autoři děkují Alejandře Vidal, Marii Konoshenko, Ilyovi Gruntovovi a Jamthô Suyá za odpovědi na dotazy ohledně konkrétních jazyků. —*Andrej Nikulin, Milena Venevová*

Redakce: Ivan Deržanski (techn. red.), Hugh Dobbs, Stanislav Gurevič, Boris Iomdin, Eimear McKnight, Andrej Nikulin (odp. red.), Aleksejs Peguševs, Jan Petr, Aleksandr Piperski, Marija Rubinštejnová, Milena Venevová, Elisija Warnerová.

Český text: Jan Petr.

Hodně štěstí!