Anton Perdih

HIERARHIČNO GRUPIRANJE PODATKOVO POGOSTOSTI GLASOV V STARIH JEZIKIH

Summary

HIERARCHICAL CLUSTERING OF DATA
ON SOUND FREQUENCIES IN ANCIENT LANGUAGES

The method of Hierarchical Clustering of particular languages based on the sound frequency in them indicates that there are the closest different ways of reading of the same texts. As expected, regarding the sound frequencies are quite close to one another the language pairs Old Slovene and Old Church Slavonic; Tocharic A and B; Estonian and Finnish. Close to one another are also Old Phrygian and Messapic. Surprisingly form by this criterion the same branch as different languages as there are Basque, Greek, Venetian, Estonian, Finnish, and Mycenaean. One reason for such a result could be the missing of some other languages. On the other hand, Old Slovene and the antique Venetic form a separate branch as if they derived of the same origin. They are on the same super-branch as the Tocharian A and B. A possible explanation for this indicates the DNK Genealogy presenting data that a lot of R1a people, to which belonged also the Tocharians, emigrated about 4000 years ago from Central Europe and Danube area across the Carpathian Mountains to the east and expanded till the East Asia. Italic languages like Latin, Oscan and Umbrian form a separate branch. Etruscan and Rhaetic form a separate, distant branch which seems to have the same origin as Hittite and Luvian.

Uvod

Vsebine napisov v nekaterih starih jezikih, kot so to npr. venetski, retijski, etruščanski, frigijski, mesapski, je težko primerjati med seboj in z drugimi jeziki. Obseg besedil v teh napisih je majhen in zato ni mogoče narediti velikih baz podatkov. Celo glasovne vrednosti vseh znakov niso zanesljivo znane, nekatere so le predpostavljene. Napisi so pogosto pisani zvezno in delitev na besede ni naznačena ter jo največkrat lahko le ugibamo. Zato tudi besedišče, slovnica in druge lastnosti jezika ter njegov razvoj še niso zadovoljivo znani in jih razni proučevalci različno razlagajo. To onemogoča uporabo pristopov, ki jih uporabljajo drugi raziskovalci pri raziskavah dobro znanih jezikov. Ugotavljati ujemanje slovnične zgradbe in jezikovnega gradiva, ki to nosi, je pri nekaterih starih jezikih zaradi majhnega obsega in poškodb napisov, ki so pisani zvezno, v narečjih in z mnogimi okrajšavami, brez njihovega dobrega razumevanja, dvomljivo. Prav pri venetskih, retijskih in frigijskih napisih so zaradi teh razlogov ustreznejše primerjave pogostosti glasov.

V ta namen sta Silvestri in Tomezzoli v letih 2005 in 2007 [1, 2] uporabila pristop z ugotavljanjem jezikoslovne razdalje na podlagi pogostosti glasov. Pri tem sta prišla do presenetljive ugotovitve, da so po pogostosti glasov venetski in retijski napisi bliže pred-Trubarjevi slovenščini (v nadaljevanju stari slovenščini) kot pa jeziku njihovega časa, latinščini. Njuni rezultati so nas spodbudili, da smo podobne primerjave naredili z različnimi načini branja teh napisov, pa tudi z drugimi jeziki.

Enodimenzionalne in večdimenzionalne analize pogostosti glasov v 16 jezikih, večinoma starih, kjer je pri nekaterih od njih vprašljiva še delitev zveznega besedila na besede, potrjujejo prejšnjo ugotovitev, da sta po pogostosti glasov venetščina in retijščina bliže stari slovenščini kot starim italskim jezikom latinščini, oskijščini in umbrijščini. Po teh lastnostih sta venetščini in retijščini blizu tudi stara frigijščina in etruščanščina. Zanimiva je po tem kriteriju podobnost estonščine odnosno finščine z večino od teh jezikov. Latinščina, oskijščina in umbrijščina tvorijo poseben skupek, ki je ločen od skupka, ki ga tvorijo etruščanščina, retijščina in venetščina. Medtem ko je etruščanščina blizu retijščini, stari slovenščini, venetščini, itd, pa ni blizu hetitščini in luvijščini, iz katerih naj bi po nekaterih domnevah izhajala. Sedanja benečanščina je po pogostosti glasov skoraj enako blizu tako latinščini kot venetščini in stari slovenščini, ter ima, čeprav jo štejejo med kentumske jezike, mnogo satemskih prvin, kar daje slutiti, da so glasovne korenine zelo obstojne, in nam lahko nudijo vpogled v izvore jezikov.

Analize pogostosti glasov in njihovih kombinacij v raznih jezikih dajejo rezultate, ki bi lahko neodvisno dopolnjevali tisto vedenje o jezikih, ki izhaja iz ujemanja slovnične zgradbe in jezikovnega gradiva, ki to nosi [3].

Primerjanih je bilo 6 metod za oceno jezikovne razdalje med 17 jeziki, z enajstimi različnimi branji nekaterih starih jezikov. Različne metode so dale različne numerične vrednosti, zato je bilo treba primerjati njihove normalizirane in sortirane rezultate. Primerjave kažejo, da je starim jezikom, kot so to etruščanščina, stara frigijščina, retijščina in venetščina, po uporabljenih metodah za oceno jezikovne razdalje stara slovenščina večinoma bliže kot latinščina in grščina. Slovenščino torej lahko upravičeno uporabimo za razvozlavanje nekaterih starih jezikov [4].

Na podlagi analize pogostosti glasov v 17 jezikih so ugotovljene meje, nad katerimi je velikost baze podatkov dovolj velika, da njena velikost ne vpliva več bistveno na rezultate izvedene iz pogostosti glasov, njihovih parov in trojčkov. Te meje so: več kot 700 posameznih glasov; več kot 8000 parov glasov; več kot 30.000 trojčkov glasov. Kriteriju za posamezne glasove ustrezajo vse uporabljene baze podatkov. Kriteriju za pare glasov ne ustrezajo baze podatkov za mesapski, oskijski, starofrigijski, retijski in venetski jezik. Kriteriju za trojčke glasov ne ustrezajo tu uporabljene baze podatkov za etruščanski, hetitski, luvijski, mikenski, oskijski, starofrigijski, retijski, staroslovenski, umbrijski in venetski jezik. Zato so pri teh jezikih uporabni predvsem rezultati na podlagi pogostosti posameznih glasov. Selektivnost pristopa pa narašča v smeri: posamezni glasovi < pari glasov < trojčki glasov.

Na podlagi analize pogostosti glasov se kaže, da mikenska pisava Linear B in morebiti tudi luvijska pisava še nista dovolj dobro razvozlani in da bi bilo dobro pri njunem razvozlavanju upoštevati tudi slovanske pare glasov tipa soglasnik-soglasnik ter trojčke

glasov tipa soglasnik-soglasnik-samoglasnik in soglasnik-soglasnik-soglasnik [5]. Videti je, da pogostost glasov nakazuje substratni jezik izpred tisočletij.

K dosedanjim pristopom dodajam še pristop s hierarhičnim grupiranjem podatkov o pogostosti glasov v starih jezikih.

Podatki

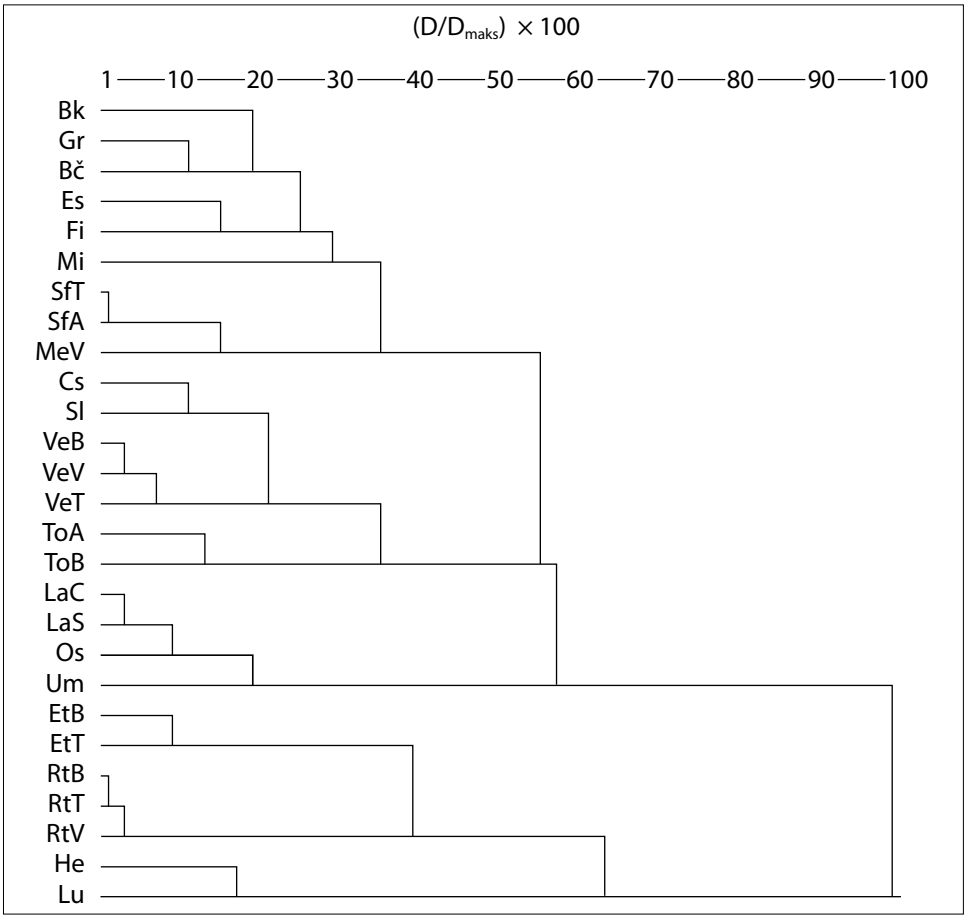
Pregled upoštevanih jezikov in načinov branja napisov je podan v tabeli 1.

Tabela 1: Uporabljeni jeziki, okrajšave, število posameznih glasov v bazah podatkov za navedene jezike [3, 6, 7]

Jezik – način branja	**Okrajšava**	**Št. glasov**
Baskovski	**Bk**	160.177
Benečanski	**Bč**	320.794
Starocerkvenoslovanski	**Cs**	458.319
Estonski	**Es**	90.742
Etruščanski – Bor; zah. jezikoslovci	**EtB, EtT**	30.421
Finski	**Fi**	449.075
Grški (Homer)	**Gr**	117.109
Hetitski	**He**	14.001
Latinski klasično	**LaC**	1.029.312
Latinski semiklasično	**LaS**	1.019.977
Luvijski	**Lu**	32.626
Mesapski	**MeV**	6.510
Mikenski	**Mi**	26.330
Oskijski	**Os**	3.057
Retijski – Bor	**RtB**	2.102
Retijski – zah. jezikoslovci	**RtT**	1.948
Retijski – Vodopivec	**RtV**	2.097
Starofrigijski – Ambrožič	**SfA**	2.290
Starofrigijski – zah. jezikoslovci	**SfT**	2.242
Stari slovenski	**Sl**	19.834
Toharski A	**ToA**	47.460
Toharski B	**ToB**	13.273
Umbrijski	**Um**	25.063
Venetski – Bor	**VeB**	7.651
Venetski – zahodni jezikoslovci	**VeT**	7.427
Venetski – Vodopivec	**VeV**	7.113

Rezultat

Rezultat uporabe metode hierarhičnega grupiranja [9] je prikazan na sliki 1.



Slika 1. Drevo upoštevanih jezikov glede na pogostost posameznih glasov

Razprava

Tako kot druge metode [3, 4, 5], tudi metoda hierarhičnega grupiranja posameznih jezikov glede na pogostost posameznih glasov v njih pokaže, da so si najbliže ista besedila, čeprav jih različni avtorji berejo nekoliko različno.

Med različnimi jeziki je bilo pričakovano, da so si po pogostosti glasov dokaj blizu pari stara slovenščina in stara cerkvena slovenščina, nadalje toharščini A in B ter estonščina in finščina.

Zanimivo je, da sta si po pogostosti posameznih glasov blizu stara frigijščina in mesapščina. Presenetljivo pa je, da pride po pogostosti posameznih glasov na isto vejo skupek baskovščina, Homerjeva grščina, benečanščina, estonščina, finščina in mikenščina.

Del vzrokov za tak rezultat je verjetno to, da niso bili upoštevani še drugi jeziki.

Stara slovenščina in venetščina sta na isti veji, kot da bi bili isto-izvorni. Skupaj sta na isti veji kot toharščini. Razlago za to nam nakazuje DNK-rodoslovje [8], ki kaže, da so se R1a ljudje, med katere spadajo tudi Toharci, pred okoli 4000 leti izselili iz Srednje Evrope in Podonavja čez Karpate na vzhod in se razširili do vzhodne Azije.

Italski jeziki latinščina, oskijščina in umbrijščina tvorijo svojo vejo, ki je po pogostosti glasov precej oddaljena od doslej omenjenih.

Etruščanščina in retijščina sta na svoji veji, ki je precej oddaljena, a videti isto-izvorna kot veja, na kateri sta hetitščina in luvijščina.

Literatura

1. M. Silvestri, G. Tomezzoli, Linguistic Computational Analysis to Measure the Distances between Ancient Venetic, Latin and Slovenian Languages. Zbornik tretje mednarodne konference Staroselci v Evropi, Založništvo Jutro, Ljubljana 2005, 77-85.
2. M. Silvestri, G. Tomezzoli, Linguistic Distances between Rhaetian, Venetic, Latin and Slovenian Languages. Zbornik pete mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2007, 184-190.
3. A. Perdih, G. Tomezzoli, V. Vodopivec, Comparison of Contemporary and Ancient Languages. Zbornik šeste mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2008, 40-87.
4. A. Perdih, Comparison of Some Methods of Estimation of Linguistic Distances. Zbornik osme mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2010, 78-86, http://www.korenine.si/zborniki/zbornik10/perdih_ling_comparisson.pdf
5. A. Perdih, Linguistic Analysis Based on the Frequency of Sound Pairs and Triplets. Zbornik devete mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2011, 27-48.
6. G. Tomezzoli, J. Kreutz, The Linguistic Position of the Tocharian, Zbornik devete mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2011, 67-86.
7. V. Vodopivec, Primerjava razumevanj mesapskih napisov, Zbornik devete mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2011, 87-130.
8. A. Perdih, Novi vidiki preteklosti. Zbornik desete mednarodne konference Izvor Evropejcev, Založništvo Jutro, Ljubljana 2012, 5-14.
9. J. Zupan, Kemometrija in obdelava eksperimentalnih podatkov, Kemijski inštitut; Inštitut Nove revije, Zavod za humanistiko 2009, 5.3 Podobnost med večdimenzionalnimi objekti, razdalja in matrika razdalj, 121-125; 7.4 Hierarhično grupiranje, 154-168; drevo 160.

Povzetek

Metoda hierarhičnega grupiranja posameznih jezikov glede na pogostost posameznih glasov v njih pokaže, da so si najbliže različna branja istih besedil. Pričakovano je bilo, da sta si po pogostosti glasov dokaj blizu stara slovenščina in stara cerkvena slovenščina, pa tudi toharščini (A in B) ter estonščina in finščina. Zanimivo je, da sta si po pogostosti posameznih glasov blizu stara frigijščina in mesapščina. Presenetljivo pa je, da pride po pogostosti posameznih glasov na isto vejo skupek baskovščina, grščina, benečanščina, estonščina, finščina in mikenščina. Del vzrokov za tak rezultat je verjetno to, da niso

bili upoštevani še drugi jeziki. Stara slovenščina in venetščina sta na isti veji, kot da bi bili isto-izvorni. Skupaj sta na isti veji kot toharščini. Razlago za to nam nakazuje DNK-rodoslovje, ki kaže, da so se R1a ljudje, med katere spadajo tudi Toharci, pred okoli 4000 leti izselili iz Srednje Evrope in Podonavja čez Karpate na vzhod in se razširili do vzhodne Azije. Italski jeziki latinščina, oskijščina in umbrijščina tvorijo svojo vejo, ki je po pogostosti glasov precej oddaljena od doslej omenjenih. Etruščanščina in retijščina sta na svoji veji, ki je precej oddaljena, a videti isto-izvorna kot veja, na kateri sta hetitščina in luvijščina.