A romániai magyar nyelvjárások atlasza informatizált térképlapjainak kvantitatív nyelvföldrajzi vizsgálata

Vargha Fruzsina Sára

1. Bevezetés

Lassan húsz éve, hogy fölmerült a magyar nyelvjárások kutatóinak körében az adatok számítógépes feldolgozásának, kutathatóvá tételének szükségessége (Balogh–Kiss 1992). A Bihalbocs elindulásával valóban lehetővé is vált régi atlaszaink hatékony, erőforrás-kímélő számítógépes feldolgozása, informatizálása, így téve nem csak könnyebben kezelhetővé, hanem a papír változatnál összehasonlíthatatlanul sokrétűbben felhasználhatóvá a bennük rejlő nyelvi adatokat (Vékás 2007).

Mit "tud" egy informatizált nyelvjárási adat egy (akár digitálisan előállított) hagyományos, térképre rajzolt adathoz képest? Az informatizált nyelvi adatok sokoldalúan hasznosíthatók és újrahasznosíthatók: kereshetők, tetszőleges szempontok szerint térképezhetők, más, azonos elvek szerint rögzített adatokkal integrálhatók, számszerű kimutatások készíthetők belőlük.

2005 és 2007 között egy NKFP finanszírozású projektben informatizáltuk A romániai magyar nyelvjárások atlasza (a továbbiakban RMNyA.) 5–8. köteteit (összesen 1200 térképlapot). Dolgozatomban ezen informatizált adatok alapján vizsgálom a nyelvi hasonlóság mértékét a RMNyA. kutatópontjai között, dialektometriai térképek segítségével. Az elemzéshez olyan kutatópontokat választottam, amelyek a dialektológia tankönyv nyelvjárási régiókat ismertető fejezetei (Juhász 2001), illetve Péntek János akadémiai székfoglalója (2005) alapján nyelv(járás)sziget helyzetűnek mondhatók.

2. Kvantitatív nyelvföldrajz

A kvantitatív nyelvföldrajzi vizsgálatok lényege, hogy több száz térképlap adatait vizsgáljuk, előre meghatározott szempontok szerint, számszerűsítve a kutatópontok adatai közti összefüggéseket. Nem néhány, a kutató prekoncepciójával összhangban álló térképlapot vizsgálunk tehát, hanem több száz térképlapot (így akár több százezer vagy akár millió nyelvi adatot), válogatás nélkül. Ennek értelmében a kvantitatív elemzési módszerek viszonylag objektívnek mondhatók.

A meghatározás szerint kvantitatív nyelvföldrajzi vizsgálatnak tekinthetjük a hangstatisztikai térképeket, amelyek bizonyos hangok előfordulási gyakoriságát, arányát mutatják meg az egyes kutatópontokon, illetve az egyes kutatópontok között (lásd pl. Bodó–Vargha 2008, Juhász 2011., Vargha 2007a).

Készíthetünk térképes kimutatást egy meghatározott korpuszban, a kutatói szempontok szerint kódolt jelenségek, nyelvi változók előfordulásáról is. Ilyen kimutatást készített Bodó Csanád A moldvai csángó nyelvjárás atlaszában előforduló román kölcsönszók területiségéről (Bodó 2007). Nyelvi változók kódolására és térképezésére azonban nem csak informatizált nyelvatlaszok, hanem területi szempontú szöveges adatbázisok felhasználásával is van lehetőség. Utóbbira példa A magyar nyelvjárások atlasza (a továbbiakban MNyA.) szövegfelvételei alapján tett kísérlet a suksükölés hatvanas évekbeli területi elterjedtségének vizsgálatára a Dunántúlon (Vargha 2007b).

A kvantitatív nyelvföldrajz talán legjellegzetesebb vizsgálati iránya a dialektometria, amelynek lényege (a fenti példáktól eltérően) nem egy-egy nyelvi változó elemzése, hanem egy nyelvatlasz (vagy annak legalább száz térképlapja) összes adatának kutatópontonkénti

összevetése, ahogyan ezt 2009-ben bemutattuk a Somogy–zalai nyelvatlasz és a MNyA. informatizált részének felhasználásával (Vargha–Vékás 2009). A továbbiakban ezt az elemzési módot, illetve a RMNyA. anyagából ezzel a módszerrel készült térképes elemzéseket mutatom be részletesebben.

3. Dialektometria

A "dialektometria" kifejezést Jean Séguy francia dialektológus használta először a 70-es években. Szomszédos kutatópontok megfelelő adatai közötti eltéréseket számszerűsített atlaszadatok alapján, és eszerint próbált nyelvjáráshatárokat megállapítani (Séguy 1973, idézi Chambers–Trudgill 1998).

A szintén romanista Hans Goebl már nem csak szomszédos kutatópontok adatait veti össze: egy atlasz minden kutatópontját minden más kutatópontjával egyenként egybeveti, egy hasonlósági mátrixot hozva létre. A mátrixból tehát bármely kutatópont bármely másik kutatóponthoz mért nyelvi hasonlóságának mértéke egy szám formájában kiolvasható. Ha egy szám a maximumhoz (mondjuk 100%-hoz) közeli, a hasonlóság mértéke nagy, kisebb szám pedig értelemszerűen gyengébb hasonlóságot jelez. A különböző számokhoz pedig különböző színek társíthatók a térképes megjelenítés során (Goebl 2006).

A groningeni egyetem kutatói, Nerbonne és Heeringa, Levenshtein algoritmusát használva vetnek össze nyelvjárási adatokat, így nem kutatói döntések, csoportosítások, hanem egy matematikai eljárás segítségével hozva létre egy, a kutatópontok közti nyelvi hasonlóság mértékét megmutató mátrixot (Nerbonne–Heeringa 1997, Heeringa 2004). Az általam felhasznált térképes dialektometriai kimutatások is ehhez hasonló módszerrel készültek, vagyis Levenshtein algoritmusának használatán alapulnak. Az algoritmus két, fonetikusan lejegyzett és betűláncnak felfogott adat egymáshoz képesti távolságát méri (azaz a különbözőség mértékét fejezi ki számszerűsített formában). Ezzel a módszerrel – térképlaponként haladva – páronként összevetjük egymással egy nyelvatlasz kutatópontjainak adatait, mindig megállapítva a két adat közötti különbözőség (illetve hasonlóság) mértékét, akár több száz kutatópont és térképlap esetében. Az összevetések számszerűsített végeredménye egy hasonlósági mátrix, amely megmutatja, átlagosan milyen arányban mutatnak hasonlóságot egymással az egyes kutatópontok adatai. Így bármelyik kutatópontról megállapíthatjuk, hogy adatai átlagosan mely kutatópontok adataival mutat nagyobb, és melyekkel kisebb hasonlóságot. (Saját magával értelemszerűen minden kutatópont esetében 100%-os hasonlóság áll fenn.) A RMNyA. 11 kötetéből eddig négy kötet áll rendelkezésünkre megfelelő formában ahhoz, hogy dialektometriai elemzést készíthessünk belőle (ez 1200 térképlapot jelent). A dialektometriai kutatásokban minimumnak tekintett, és a hasonló kimutatásokhoz felhasznált 100 térképlapnyi adatmennyiségnél ez nagyságrendekkel nagyobb korpuszt, és ezáltal várhatóan sokkal pontosabb elemzést jelent.

A RMNyA. különleges értéke a lejegyzés koherenciája, ugyanis az atlasz valamennyi adatát egyedüli gyűjtőként Murádin László jegyezte le. A finoman mellékjelezett adatok nagyban segíthetik az elemzéseinket, hiszen a hangárnyalatok (például a *felhő* szó első magánhangzójának lehetséges minőségei) jellemző területi megoszlást mutathatnak.

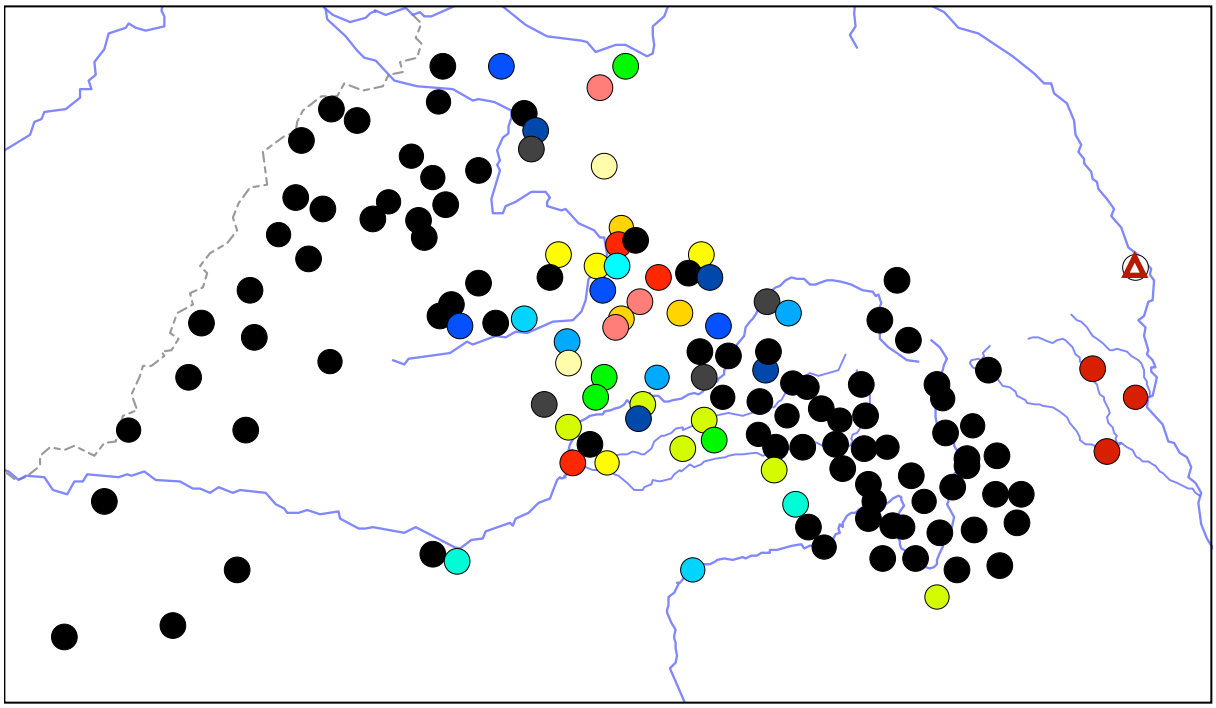
A lejegyzés informatizált változata arra is alkalmas, hogy szükség szerint konvertáljuk fonetikailag kevésbé pontos formába (eltüntetve például a mellékjeleket, akár eltekintve a magánhangzók közti minőségbeli különbségektől). A fonetikailag különböző finomságú adatok különböző arányokban mutatnak egyezéseket egymással. A *felhő,* a *fëlhő* és a *fölhő* (illetve ezek mellékjelezett vagy diftongusos változatai) például 100%-os egyezést mutatnak

egymással, ha a magánhangzók közti különbséget egy konverziós eljárás segítségével megszüntetjük. A *homáj* (és annak fonetikai variánsai) azonban még ekkor is jelentős mértékben különbözni fognak a *felhő, fëlhő,* stb adatoktól. Több hasonlósági mátrixot is készíthetünk a kutatópontjainkról aszerint, hogy milyen mértékben vesszük figyelembe a hangtani különbségeket.

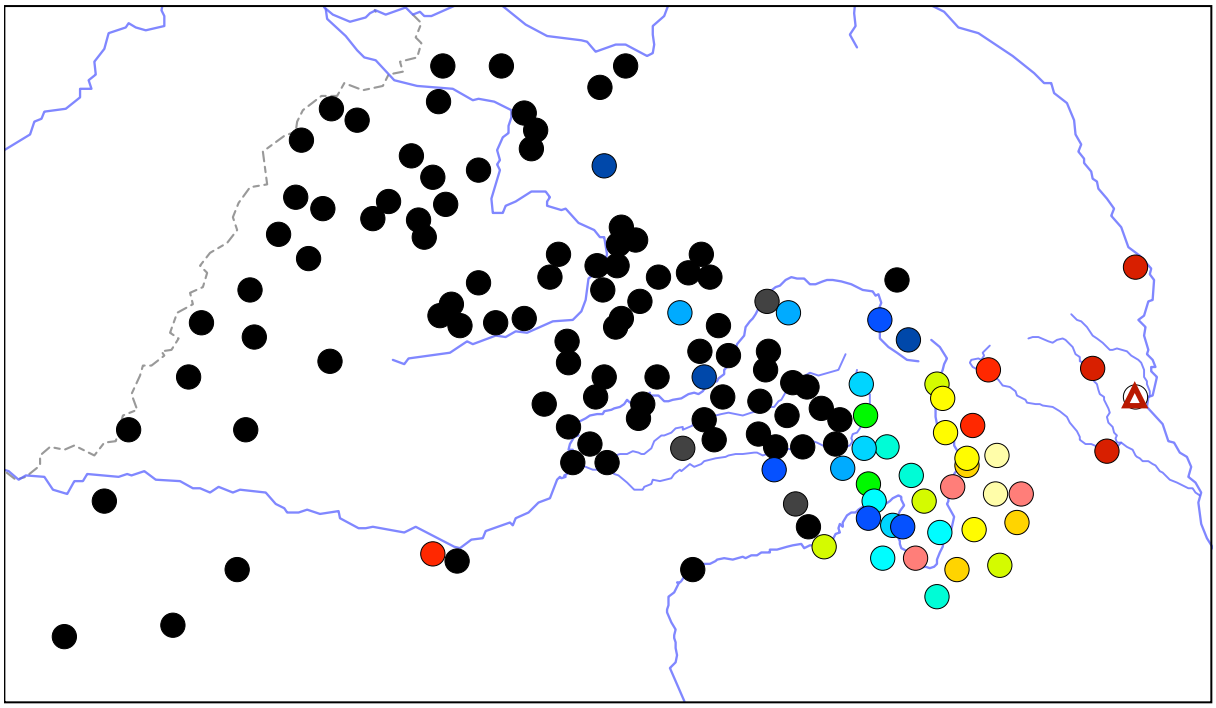
4. Dialektometriai térképek

A RMNyA. következő kutatópontjainak nyelvi hasonlósági viszonyait vizsgáltam: Szabófalva, Bogdánfalva, Pusztina, Diószeg, Csernakeresztúr, Lozsád, Oltszakadát, Halmágy, Domokos, Köröstárkány és Végvár. A térképek azt mutatják meg, hogy a kijelölt kutatóponttal mely atlaszbeli kutatópontok mutatnak leginkább nyelvi hasonlóságot (a kijelölt kutatópont földrajzi elhelyezkedését vörös háromszög jelzi). A kijelölt kutatóponthoz nyelvileg leginkább hasonló kutatópont(ok) sötét vörös színnel, a vele legkisebb mértékben hasonlóságot mutató kutatópontok fekete színnel látszanak. A két végpont között a meleg színek fokozatosan változnak át egyre hidegebb színekké, érzékeltetve ezzel az egyre kisebb fokú nyelvi hasonlóságot. Az eredményeket összegzően az 1. táblázat mutatja be.

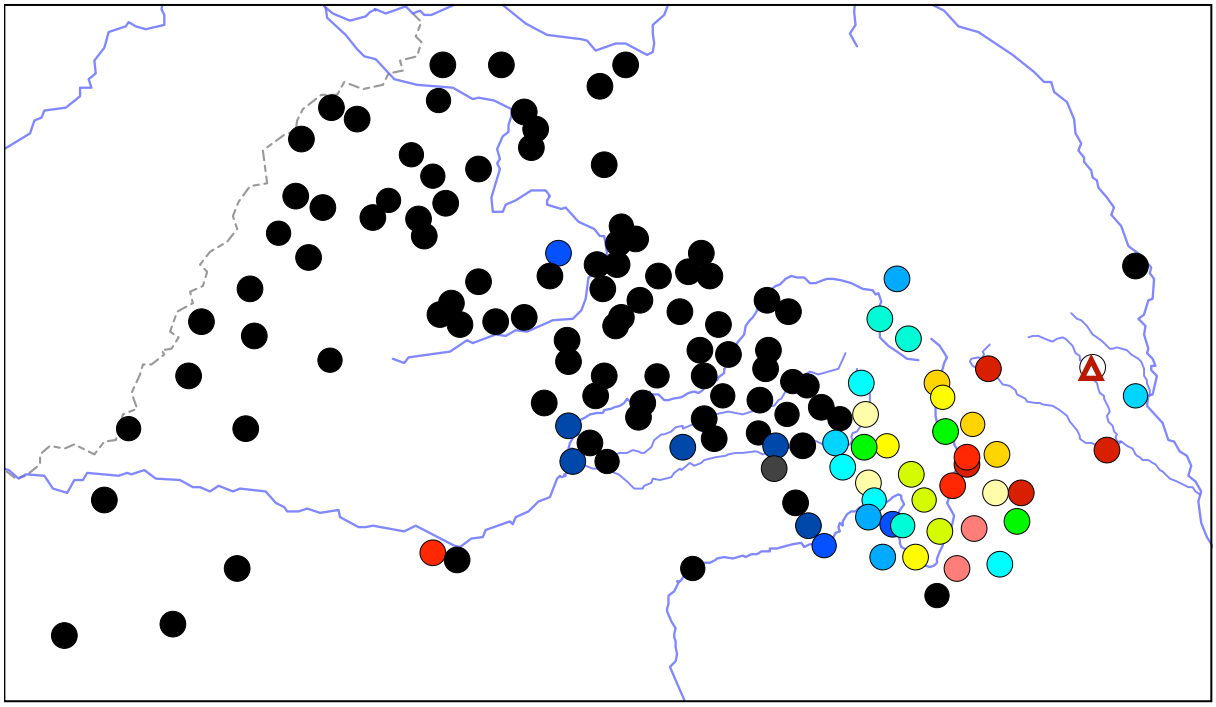
A négy moldvai kutatópont közül Szabófalva (a további három moldvai kutatóponttól markánsan eltérően) az észak-mezőségi Mezőveresegyházával és Nyíressel mutat leginkább nyelvi hasonlóságot (1. térkép). A további három moldvai kutatópont (tehát a mezőségi alapú Bogdánfalva is) leginkább keleti székely településekkel mutat hasonlóságot (2–4. térkép, a településeket név szerint lásd az 1. táblázatban). Itt jegyzem meg, hogy elemzésünk során nem szabad megfeledkeznünk arról, hogy itt csupán arra a 136 településre támaszkodhatunk, amelyek a RMNyA. gyűjtőpontjai voltak. Az itt bemutatott elemzés tehát nem mindig lehet pontos, az esetek többségében inkább irányokat, mint konkrét helyeket jelöl. Pontosabb elemzésre csak nagyobb sűrűségű kutatópont-hálózattal készült adatfelvételek, jellemzően a kisebb regionális atlaszok adatai alapján volna lehetőségünk.



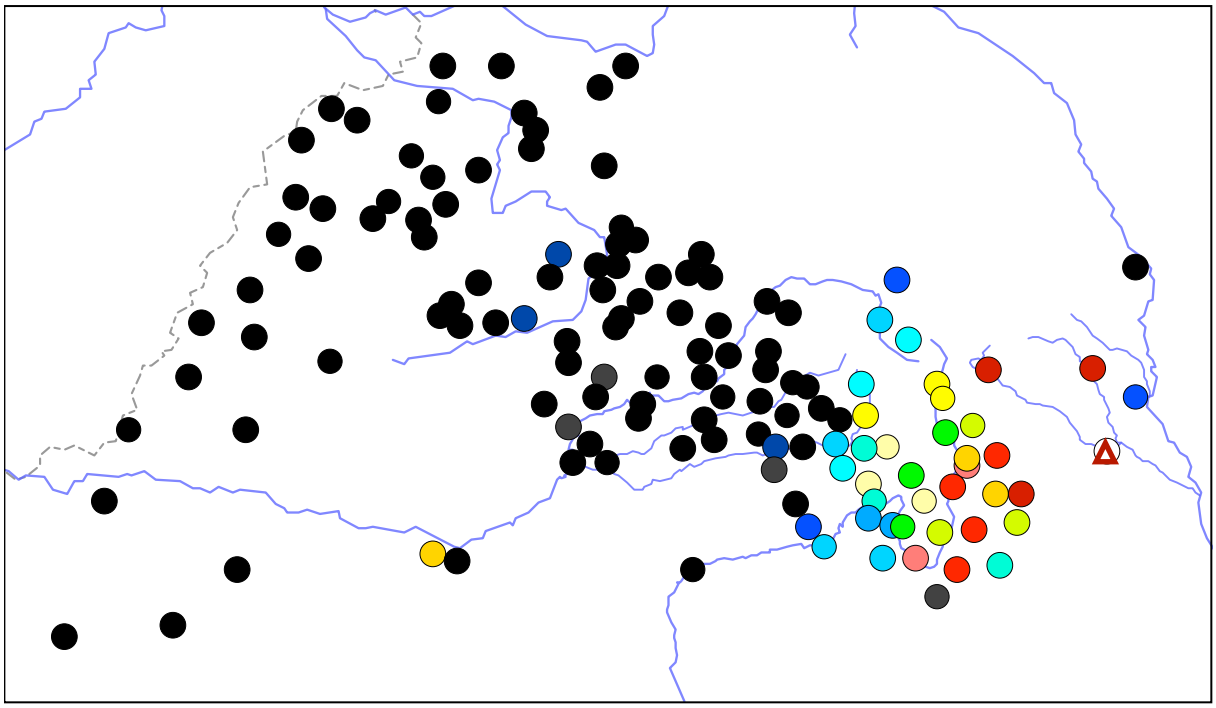
1. térkép: Szabófalva



2. térkép: Bogdánfalva

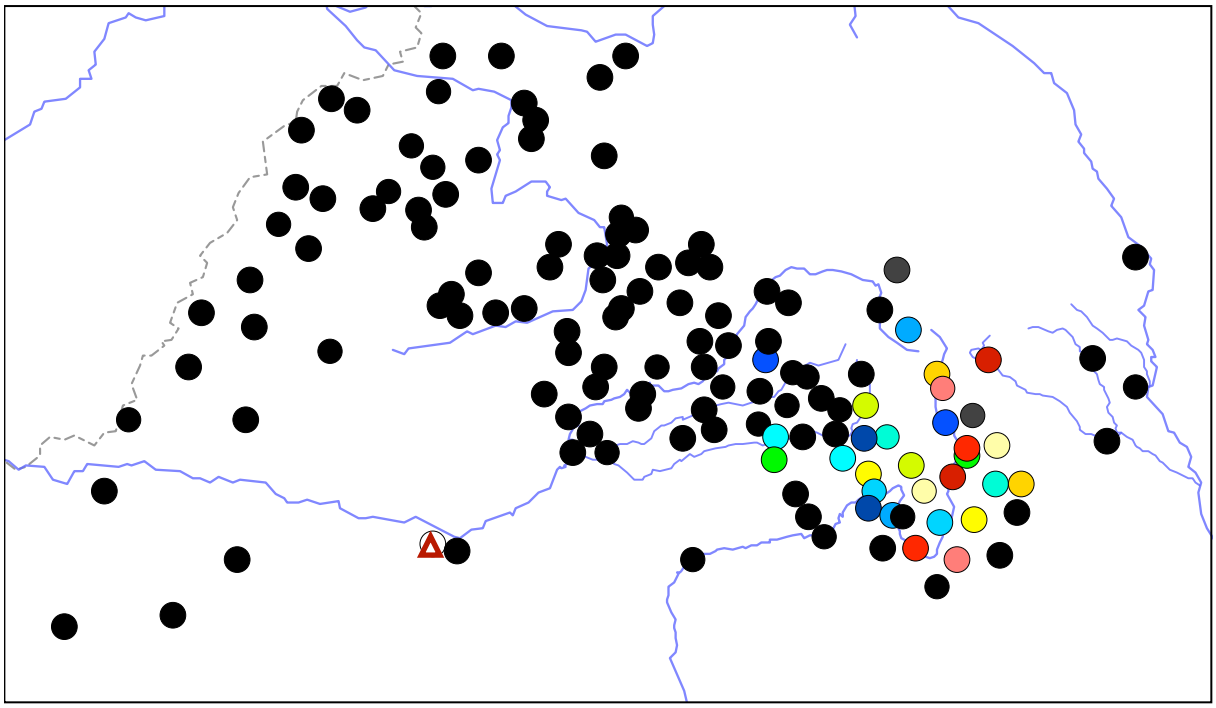


3. térkép: Pusztina

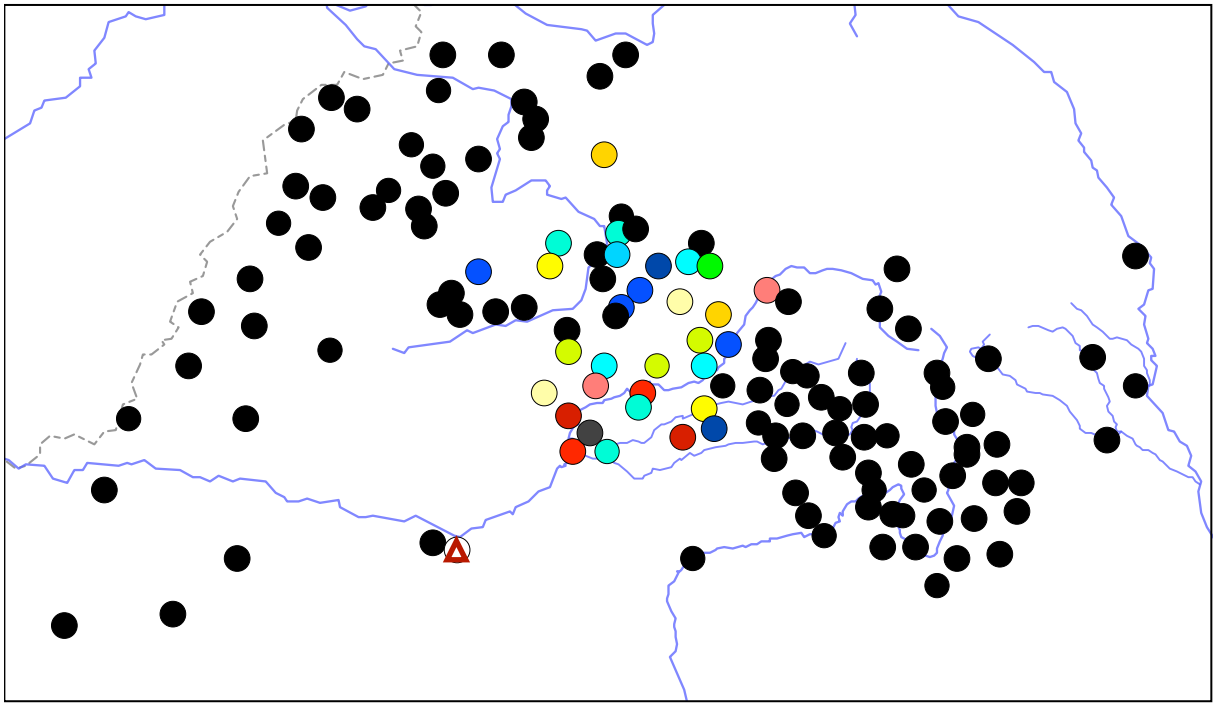


4. térkép: Diószeg

Két szomszédos kutatópont, Csernakeresztúr (5. térkép) és Lozsád (6. térkép) esetében igen látványosan eltér egymástól a nyelvi hasonlóság földrajzi súlypontja. Ha Csernakeresztúrra kattintunk az interaktív dialektometriai térképen, leginkább a Székelyföld keleti részén látunk piros színnel felvillanó, tehát nagyobb nyelvi hasonlóságot jelző kutatópontokat. Ha azonban Lozsád a kijelölt kutatópont, a Mezőség déli részén látunk melegebb színeket. Az, amit a dialektometria mutat, teljes összhangban van azzal, amit ezekről a kutatópontokról tudunk az eddigi dialektológiai kutatások alapján. Csernakeresztúr bukovinai székely település, Lozsád pedig a Közép- és Felső-Maros mentével, illetve a Küküllők vidékével mutat szorosabb nyelvi kapcsolatokat.

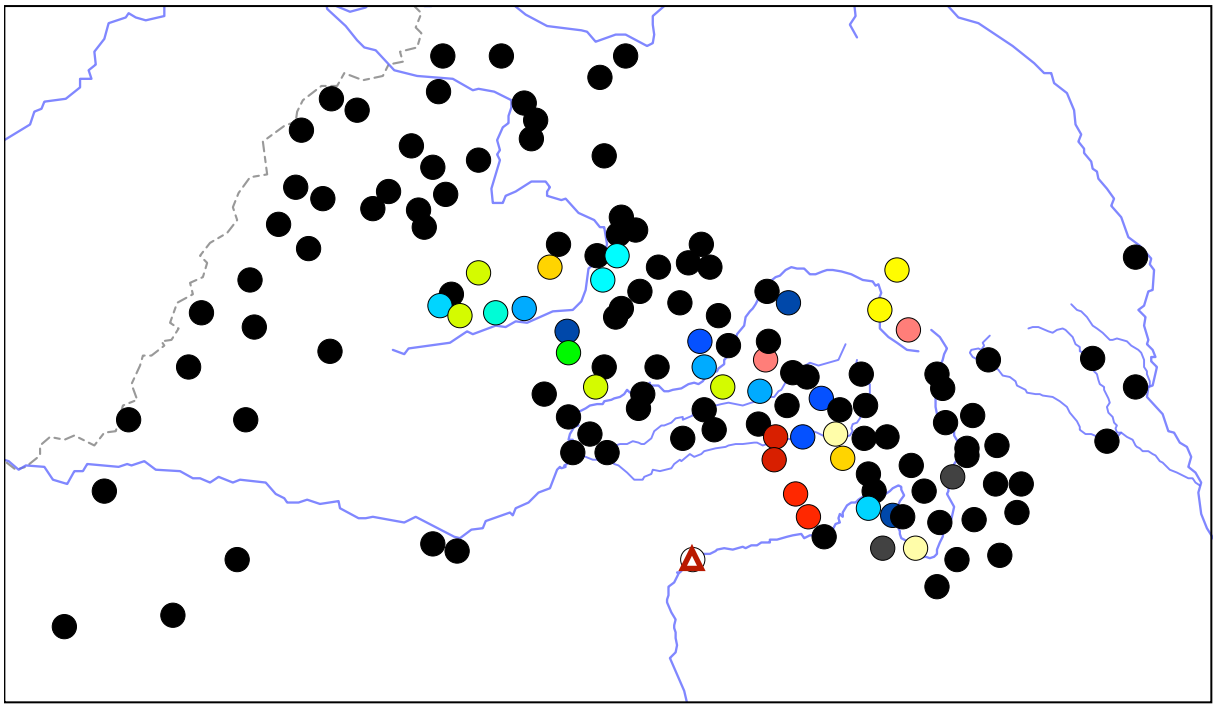


5. térkép: Csernakeresztúr



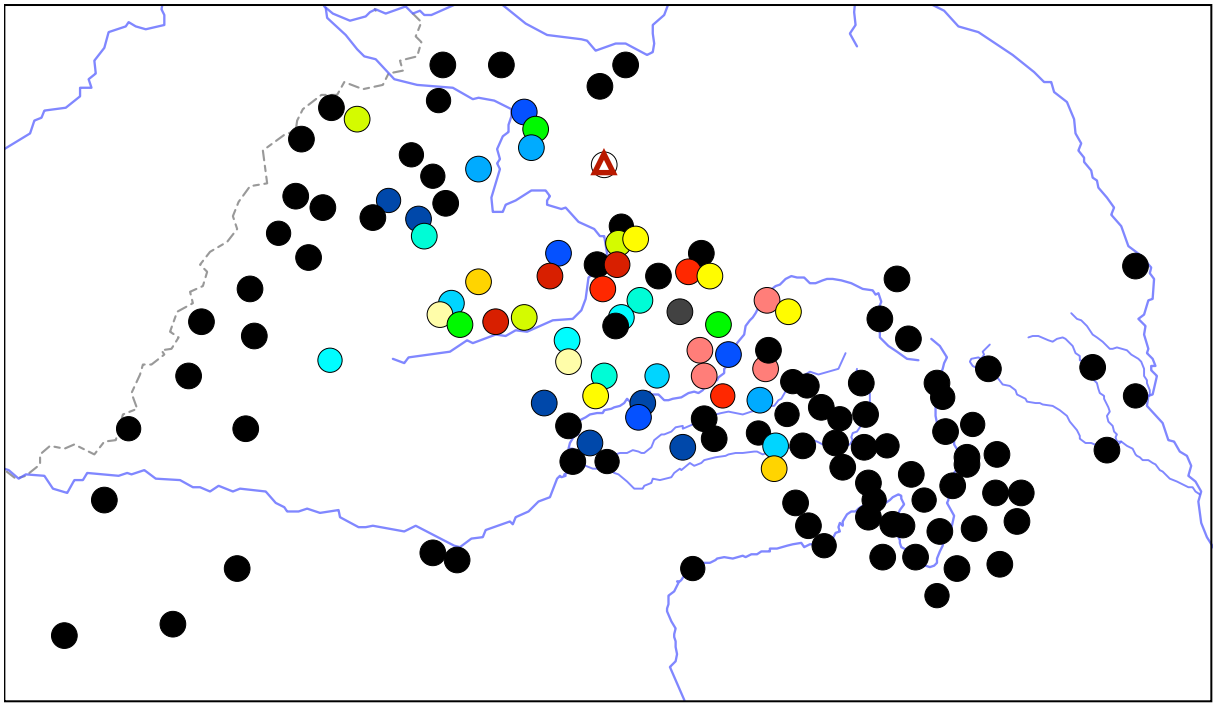
6. térkép: Lozsád

Az eddigi kutatások alapján udvarhelyszéki vonásokat mutató oltszakadáti nyelvjárás nyelvi hasonlósági súlypontja a dialektometriai térkép szerint inkább nyugatabbra esik, Búnnal és Sárpatakkal mutat leginkább egyezést az atlaszadok alapján (7. térkép).



7. térkép: Oltszakadát

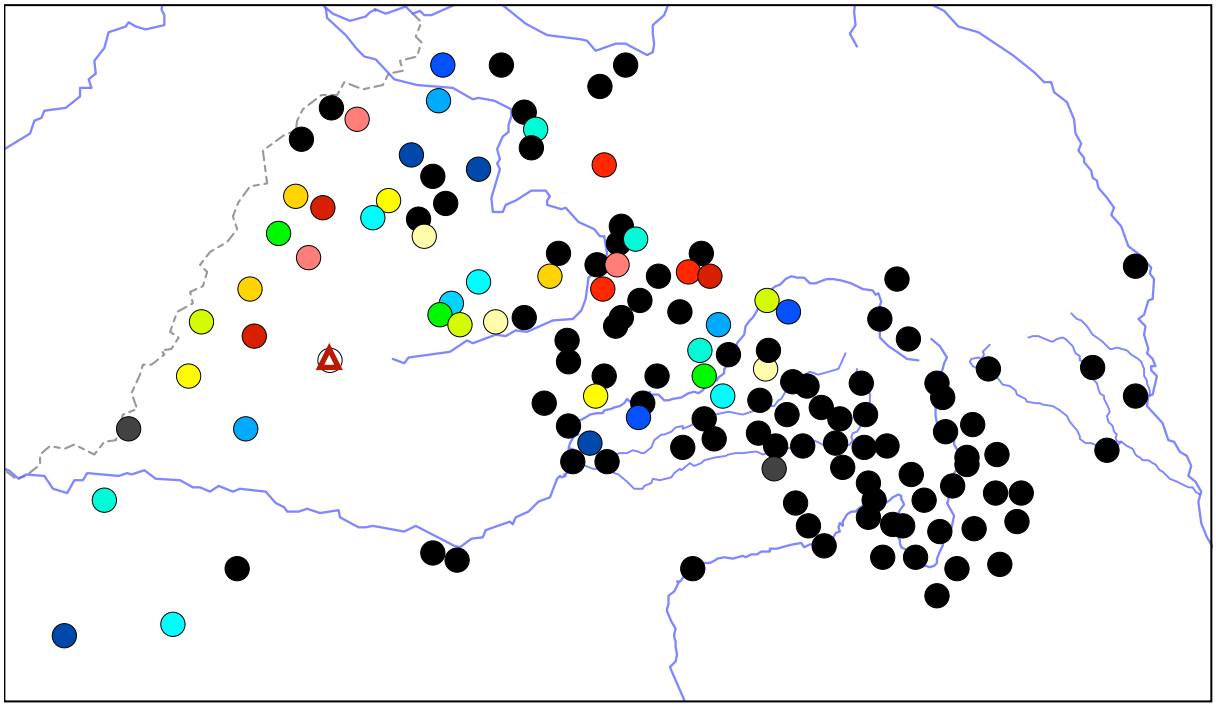
A mezőségi, mára szigethelyzetűvé vált Domokos adatai leginkább Kide és Ördöngösfüzes adataival mutatnak nagyobb nyelvi hasonlóságot, összhangban a korábbi kutatási eredményekkel. A Domokossal leginkább hasonlóságot mutató harmadik kutatópont a kalotaszegi Magyarkapus (8. térkép).



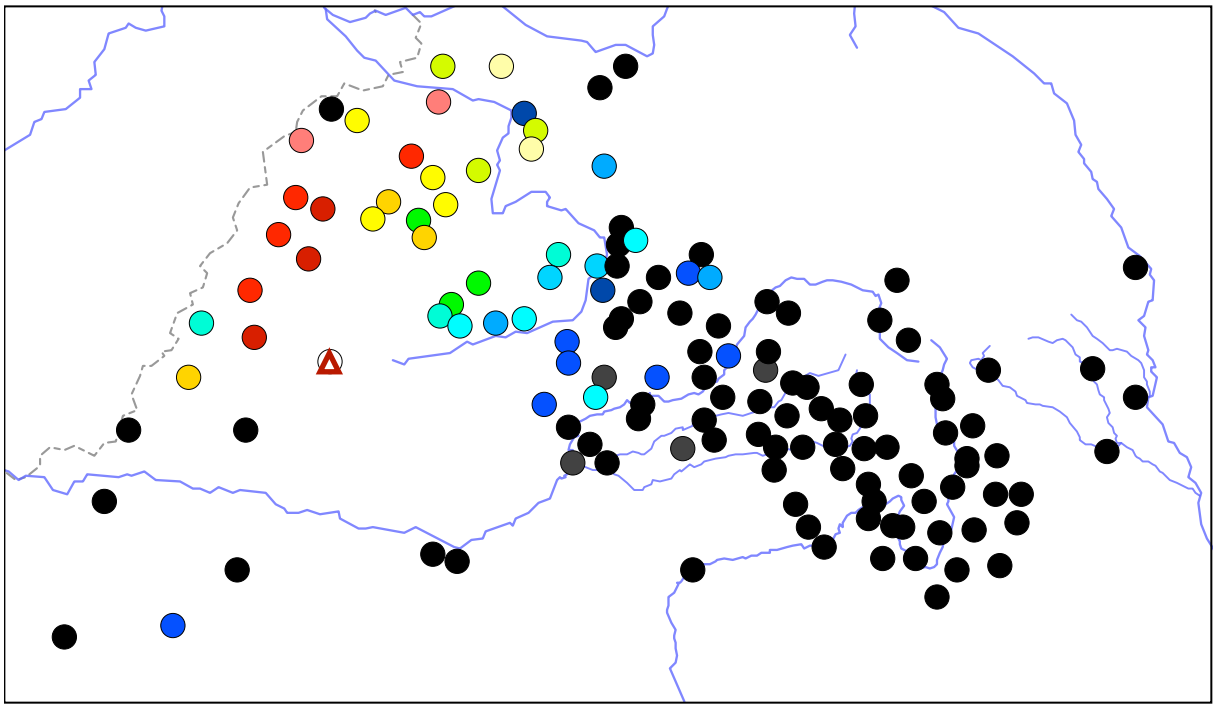
8. térkép: Domokos

Az érintkező Tisza-Körös vidéki régió hatását is mutató Köröstárkány mezőségi eredetére utal, hogy a RMNyA. kutatópontjai közül, két partiumi kutatópont mellett,

Zselykkel is erősebb nyelvi hasonlóságot mutat (9. térkép). Ha azonban megváltoztatjuk az interaktív térkép hátterében lévő mátrixot, és az apró fonetikai különbségekre is érzékeny elemzés helyett a magánhangzók minőségbeli különbségeit figyelmen kívül hagyó (lényegében a lexikai különbségekre összpontosító), fonetikailag érzéketlen mátrixot használunk, már csak a partiumi kutatópontok látszanak nagyobb nyelvi hasonlóságot jelző meleg színekkel (10. térkép).

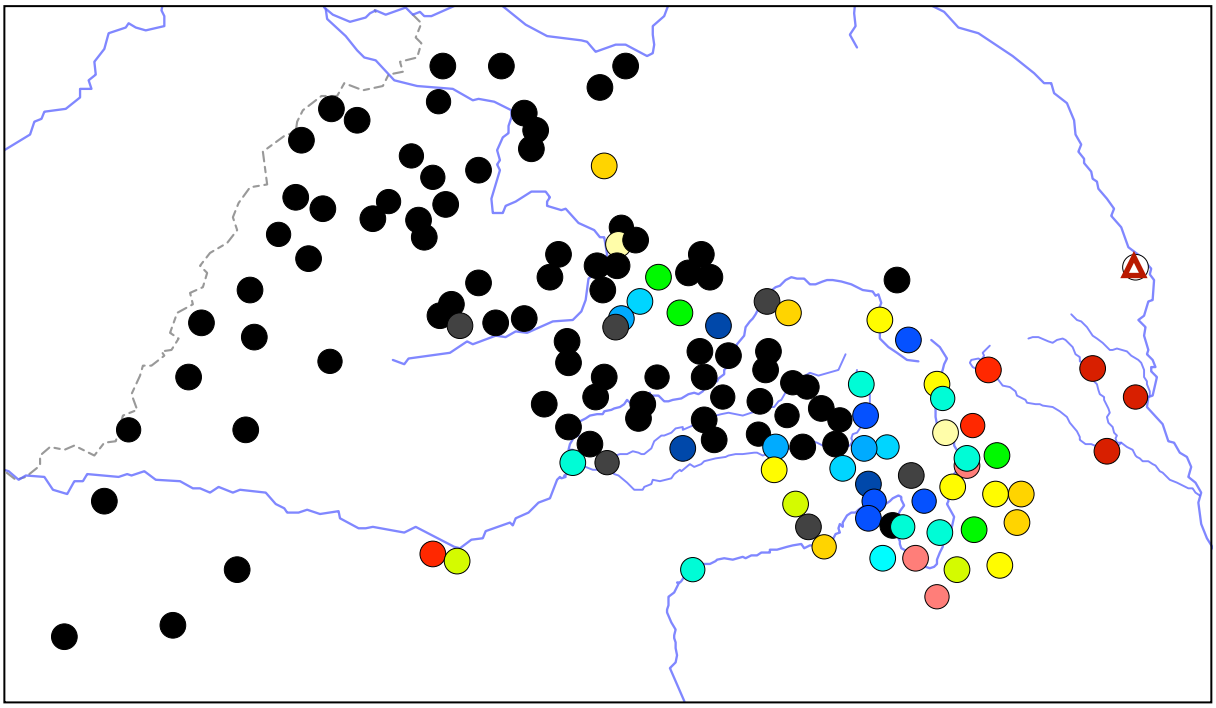


9. térkép: Köröstárkány (fonetikailag érzékeny mátrix használatával készült nyelvi hasonlósági térkép)



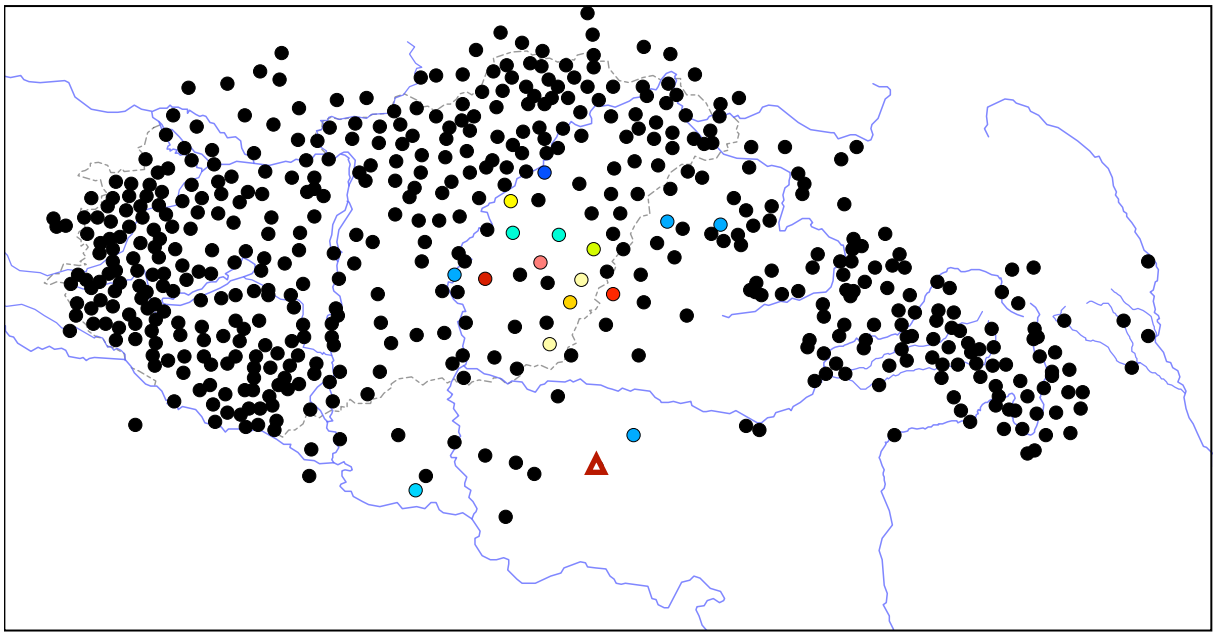
10. térkép: Köröstárkány (fonetikailag érzéketlen mátrix használatával készült nyelvi hasonlósági térkép)

A nyelvi hasonlóságot fonetikailag érzéketlen mátrix alapján vizsgálva a moldvai Szabófalva esetében is egészen megváltozik a térkép. A fonetikailag érzékeny elemzés esetén (1. térkép) meleg színekkel kirajzolódó mezőségi kutatópontok szinte egészen "eltűnnek" a látókörünkből, helyettük a térben közelebb fekvő Székelyföldön találunk melegebb színeket (11. térkép).



11. térkép: Szabófalva (fonetikailag érzéketlen mátrix használatával készült nyelvi hasonlósági térkép)

Ahogyan azt a bevezetőben már említettem, az informatizált nyelvjárási adatok, adattárak nem csupán önmagukban elemezhetők a számítógépes dialektológia módszereivel, hanem integrálhatók is egymással. A MNyA. és a RMNyA. eddig informatizált részében 164 közös, és ezért integrálható térképlap található. Ezen térképlapok adataiból integrált dialektometriai térképet készítettünk (a térkép ismeretterjesztő, internetes változata elérhető a http://www.bihalbocs.hu/dialektometria_demo.html címen). Így lehetőségünk van arra, hogy megvizsgáljuk a Tisza-Körös vidéki gyökerű Végvár nyelvi hasonlósági viszonyait. A MNyA. és a RMNyA. integrált kutatópont-hálózatában Végvár adatai legnagyobb arányban Öcsöd adataival mutatnak egyezéseket (12. térkép).



12. térkép: Végvár (A MNyA. és a RMNyA. 164 integrált térképlapja alapján)

Az elemzésbe bevont 11 RMNyA.-kutatópont feltételezett nyelvjárási kapcsolatait és a dialektometriai elemzés szerint nyelvileg leginkább hasonló kutatópontokat összefoglalóan az 1. táblázatban mutatom be. Szükséges ismét megjegyezni, hogy az általam használt korpuszban mindössze 136 romániai magyar település szerepel, így a nyelvileg hasonlónak mutatkozó települések a legtöbb esetben inkább csak irányokat, mint konkrét helyeket jelölnek számunkra.

1. táblázat: A kiválasztott kutatópontok feltételezett nyelvföldrajzi kapcsolatai a szakirodalom (Juhász 2001, Péntek 2005) alapján és a dialektometriai térképen legnagyobb nyelvi hasonlóságot mutató kutatópontok

kutatópont	eredet, kapcsolat	dialektometria (nyelvi hasonlóság)
Szabófalva	mezőségi	Mezőveresegyháza, Nyíres
Bogdánfalva	mezőségi alapú, székelyes	Gyimesbükk, Csíkmenaság
Pusztina	keleti székely	Gyimesbükk,Csíklázárfalva, Kézdialmás
Diószeg	keleti székely	Kézdialmás
Csernakeresztúr	bukovinai székely	Sepsibükszád, Gyimesbükk
Lozsád	dél-mezőségi	Magyarsáros, Miriszló
Oltszakadát	udvarhelyszéki	Bún, Sárpatak
Halmágy	székely	Zágon
Domokos	mezőségi	Kide, Ördöngösfüzes
Köröstárkány	mezőségi	Zselyk, Szék
Végvár	Tisza–Körös vidéki	Öcsöd

5. Konklúzió

Az itt bemutatott dialektometriai térképek tanulságait összefoglaló 1. táblázat alapján elmondhatjuk, hogy a dialektometria olyan, viszonylagos objektivitást is biztosító eszközt ad a szigetszerű helyzetben lévő (illetve moldvai) kutatópontok nyelvi kapcsolatainak vizsgálatához, amely sokszor határozottan megerősíti, egyben tovább pontosítja a szóban forgó beszélőközösség eredetéről vagy nyelvi kötődéseiről már rendelkezésre álló ismereteket, esetleg további elképzelések teszteléséhez mutathat irányt. A térképes

kimutatások segíthetnek tehát a nyelvjárásszigetek nyelvi kapcsolatrendszerének pontosabb feltárásában.

Noha számos esetben nagyfokú dialektometriai hasonlóságot találunk a feltételezett lakosságkibocsátó hely és a kirajzás között, fontos világosan leszögezni, hogy a földrajzi távolság ellenére megmutatkozó jelentős nyelvi hasonlóság nem értelmezhető automatikusan településtörténeti kapcsolatként.

A kiejtésbeli különbségeket különböző mértékben figyelembe vevő mátrixok feltehetően a nyelvi rendszer különböző rétegeiben mutatkozó hasonlóság kimutatására alkalmasak. A fonetikailag érzékeny elemzéssel szemben az adatok lejegyzésének "durvításán" alapuló eljárás inkább csak a lexikai különbségeket és hasonlóságokat érzékeli, így az ilyen dialektometriai térképeken a nyelv(járás)szigetek már nem a feltételezett kibocsátó hellyel (vagy annak környezetével), hanem rendszerint a hozzájuk térben legközelebb eső kutatópontokkal mutatnak nagyobb nyelvi hasonlóságot. A különböző mátrixok által kirajzolt térképek összevetése alapján úgy látszik, leginkább a fonetikailag érzékeny elemzés alkalmas a lehetséges nyelvföldrajzi kapcsolatok kimutatására térben távoli kutatópontok között. Ez a megállapítás összhangban áll azzal a feltételezéssel, hogy a nyelvjárásközi összevető vizsgálatokban a hangtannak van kitüntetett szerepe.

Az azonos eljárással, azonos kódrendszer szerint informatizált nyelvjárási adatok integrálhatók, integráltan elemezhetők. Az integrált dialektometriai térképek segítségével a különböző adattárak kutatópontjai közti nyelvi hasonlóság is vizsgálható.

Bibliográfia

Balogh Lajos – Kiss Gábor (1992). A magyar nyelvjárások atlaszának számítógépes feldolgozása. In: Kontra Miklós (szerk.). *Társadalmi és területi változatok a magyar nyelvben.* Budapest: MTA Nyelvtudományi Intézet: 5–17

Bodó Csanád (2007). A moldvai magyar nyelvjárások román kölcsönszórétegének területisége. In: Benő Attila – Fazakas Emese – Szilágyi N. Sándor (szerk.). *Nyelvek és nyelvváltozatok. Köszöntő kötet Péntek János tiszteletére. I. kötet.* Kolozsvár: Anyanyelvápolók Erdélyi Szövetsége Kiadó: 160–174.

Bodó Csanád – Vargha Fruzsina Sára (2008). Régi nyelvatlaszok – új módszerek. *Magyar Nyelv* 104: 335–351.

Chambers, Jack – Peter Trudgill (1998). *Dialectology*. Cambridge: Cambridge University Press.

Goebl, Hans (2006). Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21: 411–435.

Heeringa, Wilbert (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen Dissertations in Linguistics 46. Groningen.

Juhász Dezső (2001). A magyar nyelvjárások területi egységei. In: Kiss Jenő (szerk.). *Magyar dialektológia.* Budapest: Osiris Kiadó: 262–324.

Juhász Dezső (2011). A magyar nyitódó kettőshangzók történetéről tér és idő dimenziójában. In: Bakró-Nagy Marianne – Forgács Tamás (szerk.). *A nyelvtörténeti kutatások újabb eredményei VI.* Szeged: Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék. 123–128.

Nerbonne, John – Wilbert Heeringa. (1997). Measuring Dialect Distance Phonetically. In: John Coleman (ed.). *Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics*. Madrid: 11-18.

Péntek János (2005). Magyar nyelv- és nyelvjárásszigetek Romániában. *Magyar Nyelv* 101: 406–413.

Séguy, Jean (1973). La dialectométrie dans l'atlas linguistique de la Gascogne. *Revue de linguistique romane* 37: 1–24.

Vargha Fruzsina Sára (2007a). Állatok kicsinyeinek megnevezése a keleti magyar nyelvjárásokban. In: Hoffmann István – Juhász Dezső (szerk.). *Nyelvi identitás és a nyelv dimenziói.* Debrecen–Budapest: Nemzetközi Magyarságtudományi Társaság: 237–248.

Vargha Fruzsina Sára (2007b). Nyelvi változók a Magyar nyelvjárások atlasza hangfelvételeiben. In: Guttmann Miklós – Molnár Zoltán (szerk.). *V. Dialektológiai Szimpozion. Szombathely, 2007. augusztus 22–24.* Szombathely: Berzsenyi Dániel Főiskola: 279–288.

Vargha Fruzsina Sára – Vékás Domokos (2009). Magyar nyelvjárási adattárak vizsgálata interaktív dialektometriai térképekkel. Előadás a Magyar Nyelvtudományi Társaság felolvasóülésén. 2009. március 24. http://bihalbocs.hu/eloadas/dialektometria_20090324.pdf

Vékás Domokos (2007). Számítógépes dialektológia. In: Guttmann Miklós – Molnár Zoltán (szerk.). *V. Dialektológiai Szimpozion. Szombathely, 2007. augusztus 22–24.* Szombathely: Berzsenyi Dániel Főiskola: 289–293.