VARGHA FRUZSINA SÁRA

ATLASZINTEGRÁLÁS ÉS KVANTITATÍV ADATELEMZÉS*

1. Bevezetés. A magyar nyelvföldrajz jövőjéről, aktuális feladatairól szóló 2007-es tanulmányában JUHÁSZ DEZSŐ még elvégzendő feladatként veszi sorra legjelentősebb nyelvatlaszaink adatainak informatizálását (megfelelő számítógépes rögzítését), illetőleg a már megkezdett munkálatok befejezését, valamint az adatok egységes adatbázisba rendezését.

Két legnagyobb nyelvatlaszunk, A magyar nyelvjárások atlasza (a továbbiakban MNyA.) és A romániai magyar nyelvjárások atlasza (a továbbiakban RMNyA.) számítógépes feldolgozása azért is lényeges kérdés, mert csak integrálásukkal végezhető el a teljes magyar nyelvterület átfogó, kvantitatív dialektológiai elemzése. A magyar nyelvjárási adattárak számítógépes földolgozását elsősorban a Bihalbocs programmal végezzük, amely számtalan lehetőséget biztosít nemcsak az adatrögzítés, hanem az adatelemzés területén is (a számítógépes dialektológia alapvetéseiről lásd VÉKÁS 2007).

JUHÁSZ DEZSŐ 2007-es összefoglaló tanulmányának megjelenése óta sok minden történt: 2014-ben befejeződött a MNyA. (nyomtatásban megjelent térképlapjainak) informatizálása, a RMNyA. nyomtatásban megjelent 3297 térképlapjából 2837 informatizálva van, ebből 1516 térképlap informatizált változatának az ellenőrzése is megtörtént.[1] Dolgozatomban az adattárak integrálásának legfontosabb feltételeit, e téren elért eredményeinket és a fölmerülő problémákat mutatom be röviden, különös tekintettel a kvantitatív adatelemzésre.

2. Atlaszintegrálás. A bevezetőben már említettem, hogy a nyelvatlaszok feldolgozása során informatizált adatokat hozunk létre. Az informatizálás lényege, hogy a feldolgozás során minden, az adatra vonatkozó, rendelkezésünkre álló információt megőrzünk (adattár, kutatópont, esetleg adatközlőre vonatkozó szociológiai adatok, hangfájlrészlet, ahol az eredeti hangzó forma megtalálható).

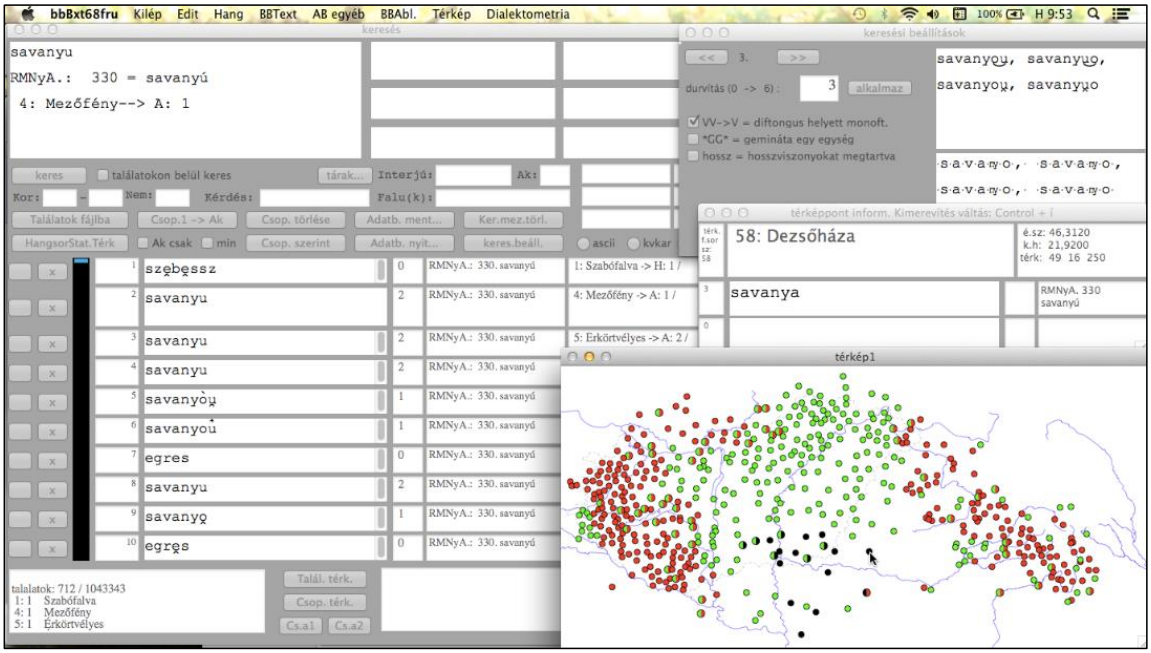
Az informatizált adatok egyik előnye, hogy újrahasznosíthatók: nemcsak egyszer, egyetlen térkép elkészítéséhez használhatjuk föl azokat, hanem többféleképpen, akár egyesítve más adattárakból származó informatizált adatokkal.

A különböző atlaszokból származó adatok integrálásához azonban feltétlenül szükségünk van integrált kutatópont-hálózatra. Két (vagy több) adattár kutatópont-hálózatának összekapcsolása természetesen nem egyszerűen csak annyit jelent, hogy az adattárakban kutatópontként szereplő valamennyi települést rárajzoljuk (automatizáltan rávetítjük) a térképre. Rendszerünknek pontosan tudnia kell, melyik pont melyik adattárnak eleme, illetve azt is, ha egyazon kutatópont egyszerre több adattárban is szerepel.

* A dolgozat a PD108442 számú OTKA-kutatás keretében készült.

[1] Az adattárak számítógépes rögzítése több lépcsőben történt, a folyamatot több projektum is támogatta. A MNyA. rögzítését Balogh Lajos indította el, még az 1990-es évek elején, a MTA Nyelvtudományi Intézetében. A MNyA. és a RMNyA. informatizálásának munkálatait jelentősen támogatta a 5/056/2004. számú NKFP-projekt (vezető kutató Kiss Jenő, a részfeladat felelőse Juhász Dezső volt), illetve a PD108442 számú OTKA-kutatás (projektvezető: Vargha Fruzsina Sára).

1. ábra
Keresés a Bihalbocsban: integrált atlaszok adatainak csoportosítása és térképezése (*savanyú* = kék, *savanyó* = piros, *savanya* = fekete)



A keresést, csoportosítást és térképezést az informatizált nyelvatlaszok integrált adatbázisában az 1. ábra szemlélteti. A példában a MNyA. 695. és a RMNyA. 330. számú, 'savanyú' címszavú térképlapjainak adataiban kerestem meg és csoportosítottam a *savanyú* különböző alakváltozatait az utolsó magánhangzó minősége szerint. A keresést az adatokban a fonetikus lejegyzés automatikus egyszerűsítése segíti, így nem kell minden egyes mellékjelezett változatot egyenként számba venni (az éppen alkalmazott beállítás tulajdonságait a jobb fölső ablak mutatja). A csoportosítás során nem feltétlenül kell valamennyi adatot figyelembe vennünk, itt például nem kaptak kódot (és így nem jelennek meg a térképen sem) az *egres* típusú adatok.

3. **Aggregált adatok elemzése.** A magyar nyelvjárási adattárak közül valamennyi a magyar egyezményes hangjelölési rendszer használatával készült, gondolhatjuk tehát, hogy a fent leírtak elegendők az adattárak integrált elemzésének biztosításához. Akkor azonban, ha az adatainkat nem önállóan szeretnénk vizsgálni, hanem kvantitatív módszerekkel aggregált adatokat szeretnénk létrehozni, újabb problémával szembesülhetünk: ugyan valamennyi adattár gyűjtői és lejegyzői a magyar egyezményes hangjelölést használták, ezt mégsem teljesen egyező gyakorlat mentén tették. Márpedig azokban a vizsgálatokban, ahol a fonetikus lejegyzés automatikus elemzésével dolgozunk, a hangjelölés alkalmazásában rejlő esetleges különbségek – megtévesztő módon – területi különbségekként jelenhetnek meg.

Jelen tanulmány keretei nem teszik lehetővé, hogy teljes részletességgel áttekintsem a lejegyzési szokások közti eltéréseket, az *ó^{u}*-féle, középső nyelvállású veláris labiális előtagú és felső nyelvállású veláris labiális utótagú diftongusok példáján mutatom be, milyen jellemző különbségek vannak a diftongusok lejegyzésében. Ilyen fajta diftongusok lejegyzésére a MNyA. 27, míg az RMNyA. 43 különböző jelölésmódot alkalmaz (természetesen igen eltérő gyakorisággal). A MNyA. esetében az öt leggyakoribb lejegyzési megoldás lefedi az összes eset (N = 8865) 97%-át, míg a RMNyA. esetében a tíz leggyakoribb jelölésmód is csak az 1516 felhasznált térképlap adataiban előforduló esetek (N = 7186) 90%-át adja ki (lásd az 1. táblázatot; a hagyományosan emeléssel jelölt, nyomaték nélküli elemet rendszerünkben a magánhangzó fölé tett, fölfelé mutató nyíl fejezi ki).

1. táblázat
A középső nyelvállású veláris labiális előtagú és felső nyelvállású veláris labiális utótagú diftongusok leggyakoribb jelölési módjai a MNyA.-ban és a RMNyA.-ban

RMNyA.	darab	MNyA.	darab
óu̯	1706	óû̯	5230
óû̯	782	o͡u̯	2045
òu̯	778	o͡û̯	613
òû̯	696	ô̯û̯	412
oû̯	605	ô̯û̯	298
ō͡u̯	566	òû̯	91
ō͡û̯	477	oû̯	41
ou̯	396	ô̯û̯	27
ò̯u̯	240	ò̯û̯	25
ó̯u̯	210	óû̯	17

Nyilvánvaló tehát, hogy a RMNyA.-ban sokkal differenciáltabb a diftongusok jelölése. Ha megnézzük, melyek a leggyakoribb jelölésmódok a két atlaszban, egyértelművé válik, hogy a gyakorlat igen eltérő. A MNyA.-ban második leggyakoribb jelölésmód a RMNyA.-ban csak elvétve jelenik meg, be sem kerülhetett az első tízbe. A RMNyA.-ban leggyakrabban használt jelölésmód pedig egyáltalán nem szerepel a MNyA.-ban, hiszen ez utóbbi adattár csak két fokozatot különít el a diftongusok jelölésére a nyomatékeloszlás szempontjából.

Az itt szemléltetett példa más diftongusok jelölését is jellemzi, és természetesen vannak további különbségek is. A RMNyA.-ban például az úgynevezett széles ejtésű, a szokványosnál kissé nyíltabban ejtett hosszú magánhangzókat hosszúságjellel ellátott rövid megfelelőjük jelöli (pl. *ō*, *ő̄*). A MNyA. hangjelölési gyakorlatára ez a jelölésmód nem

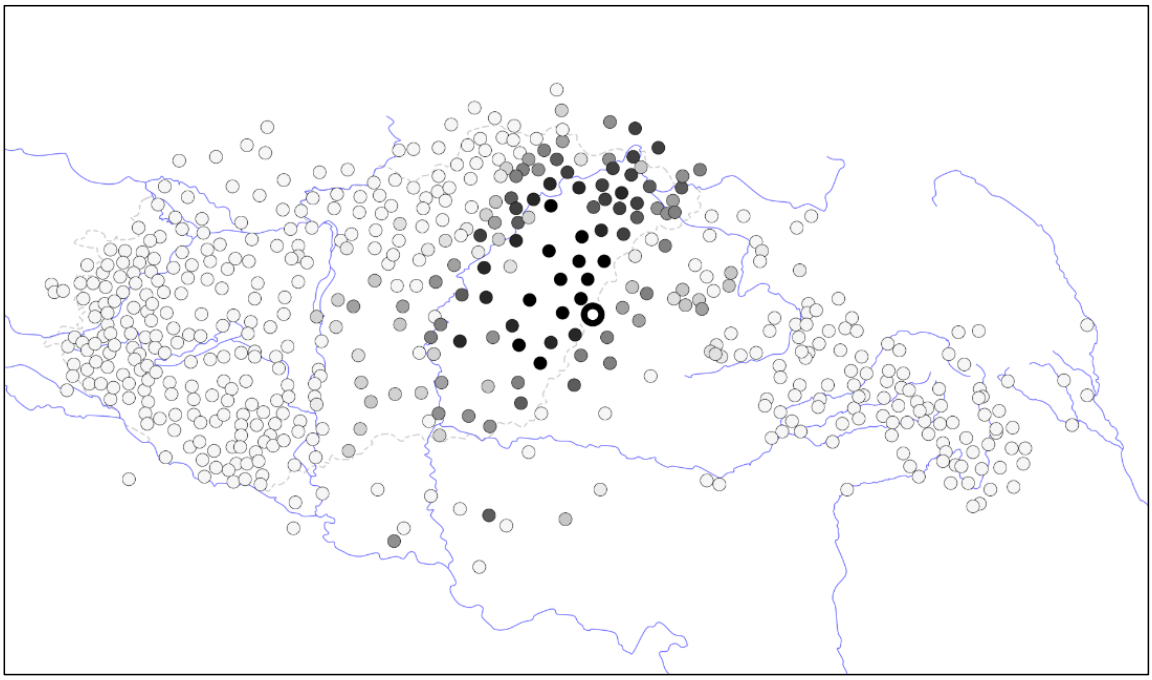
jellemző, ott inkább a hosszú hangok nyíltabb mellékjeles változataival találkozhatunk, feltehetőleg hasonló funkcióval.

A kvantitatív nyelvföldrajzi kutatások körébe tartozó dialektometria lényege nem egy-egy nyelvi változó elemzése, hanem egy nyelvatlasz (vagy annak legalább száz–kétszáz térképlapja) összes adatának kutatópontonkénti összevetése. HANS GOEBL és munkatársai a térképlapok adatainak osztályozásával, csoportosításával végeztek dialektometriai vizsgálatokat nyelvatlaszadatokon (GOEBL 2002, 2005, 2006). A groningeni egyetemen JOHN NERBONNE és munkatársai a fonetikusan lejegyzett adatokat automatikusan, Levenshtein algoritmusának alkalmazásával vetik össze, így számszerűsítve a nyelvi adatok hasonlóságának mértékét (HEERINGA 2004, NERBONNE–HEERINGA 2013). A hasonlóság 100% két teljesen azonos adat esetén, és 0% akkor, ha egyetlen közös hang sincs bennük (a *savanyú* és a *savanya* például majdnem 100%-os, míg a *savanyú* és az *egres* minimális, közel 0%-os hasonlóságot mutat).

Az összevetések számszerűsített végeredménye egy hasonlósági mátrix, amely megmutatja, átlagosan milyen arányban mutatnak egyezést egymással az egyes kutatópontok adatai. Így bármelyik kutatópontról megállapíthatjuk, hogy adatai átlagosan mely kutatópontok adataival mutatnak nagyobb, és melyekkel kisebb hasonlóságot. Az itt látható térképek hátterében lévő hasonlósági mátrixok a groningeni módszer alapján, vagyis az adatok Levenshtein-alapú automatikus összevetésével készültek (a módszer magyar nyelvatlaszokon történő alkalmazására lásd még VARGHA–VÉKÁS 2009, VARGHA 2010, VARGHA megjelenés előtt).

2. ábra

Ártánd (MNyA.) nyelvi hasonlósági térképe a MNyA. és a RMNyA. 482 integrált térképe alapján (eredeti mellékjeles lejegyzés)



A 2. és a 3. ábra Ártánd nyelvi hasonlósági térképét mutatja a MNyA. és a RMNyA. 482–482 térképlapjának integrált dialektometriai elemzése alapján. A kijelölt kutatóponthoz (vastag fekete körvonallal) viszonyítva fekete, ill. sötétszürke színek jelzik a nagyobb nyelvi hasonlóságot. A 2. ábra az eredeti, mellékjeles lejegyzés, a 3. ábra a mellékjeleket nem tartalmazó, módosított lejegyzés alapján készült.

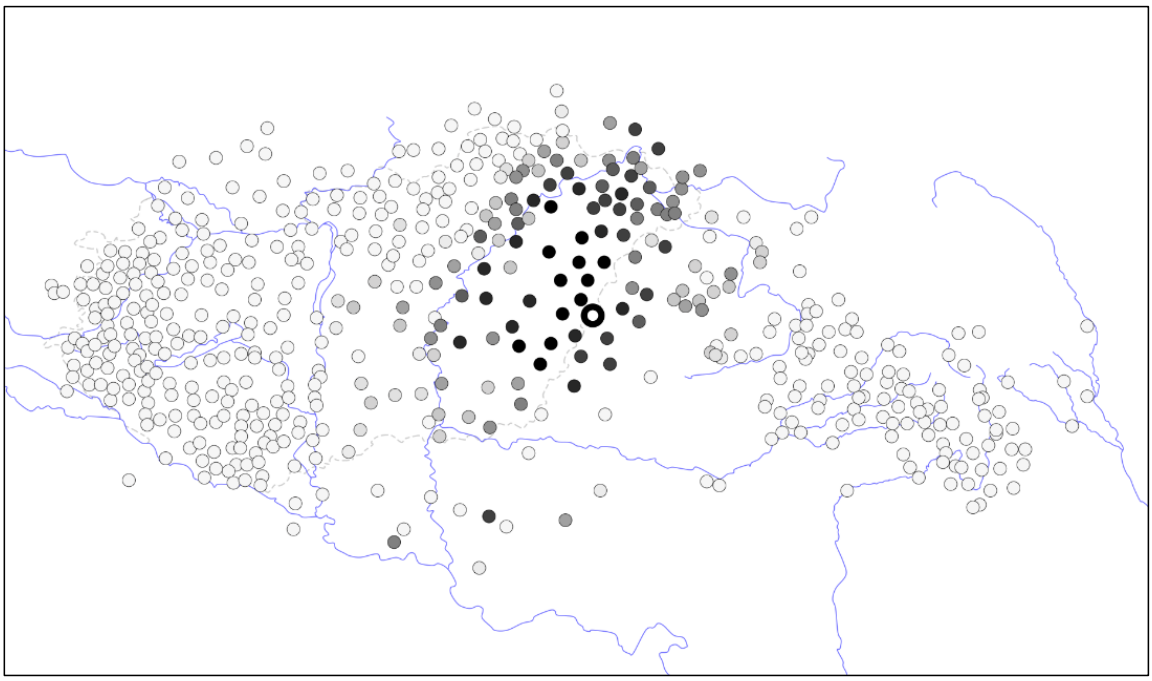
Ártándon jellemző a záródó diftongusok használata, ami (feltehetőleg nem kizárólagosan) felelős lehet azért, hogy a 2. ábrán az országhatár, amely esetünkben a két integrált adattárat is egymástól földrajzilag elhatárolja, nyelvjárási törésvonalnak tűnhet, a nyelvi hasonlósági súlypont legalábbis erőteljesen nyugatra mutat. Amennyiben eltekintünk a mellékjelek használatától, vagyis automatikusan leegyszerűsítjük a lejegyzést, megszüntetve a diftongusok nyomatékeloszlásának jelölésében rejlő különbségeket is, az adattárak közti határ érzékelhetően kevésbé markánsan jelentkezik (a 3. ábrán az Ártándtól keletre fekvő kutatópontok sötétebbek lesznek, nagyobb nyelvi hasonlóságot jelezve).

Föltehető, hogy sokkal koherensebb integrált elemzést készíthetnénk olyan eljárás kidolgozásával, amely pontosan figyelembe veszi, és módszeresen megszünteti a hangjelölési gyakorlat eltéréseit. A mellékjelek automatikus kivonása az adatokból már előrelépést jelent, valószínűsíthető azonban, hogy nem szüntet meg minden, lejegyzési szokásokkal összefüggő különbséget, míg esetleg nyelvileg pertinens részleteket eltüntethet.

Az aggregált adatokon végzett további elemzések szempontjából elengedhetetlen, hogy a dialektometriai eljárással készített mátrix valóban releváns nyelvi hasonlósági értékeket mutasson. Ilyen eljárásra mutat példát a 4. ábra, ahol a RMNyA. 1516 térképlapjának adatai alapján készített nyelvi hasonlósági mátrix adataiból többdimenziós skálázással készült térkép látható.

3. ábra

Ártánd (MNyA.) nyelvi hasonlósági térképe a MNyA. és a RMNyA. 482 integrált térképe alapján (mellékjelek nélküli lejegyzés)

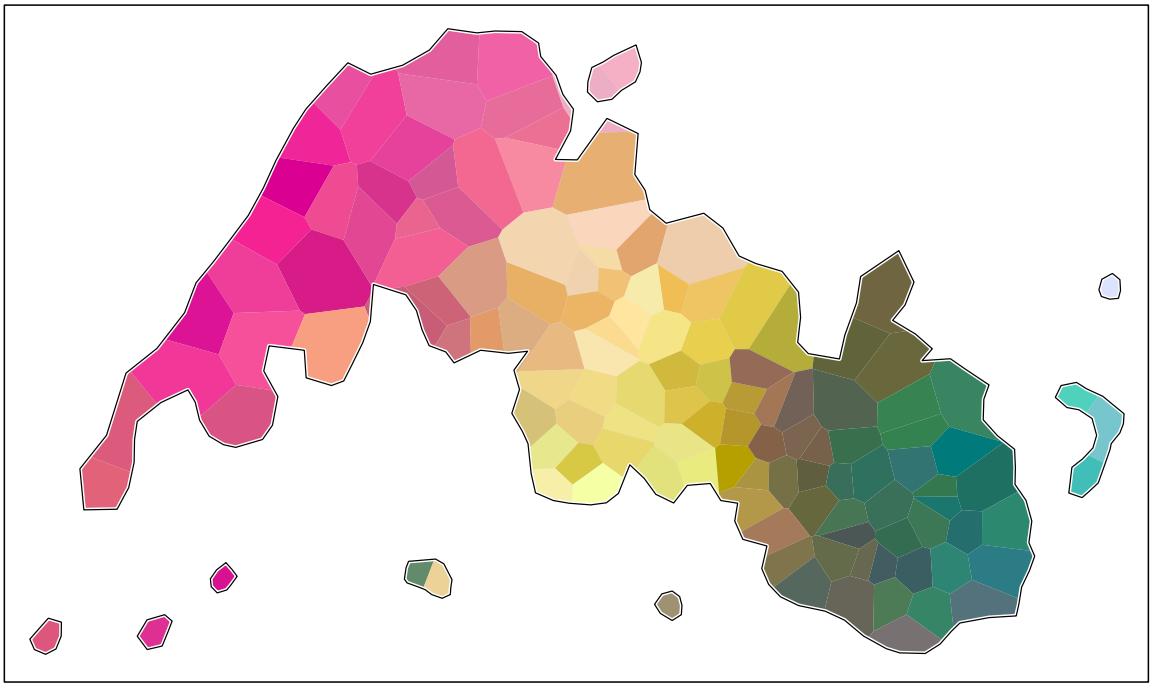


A többdimenziós skálázás jelenleg az egyik legmegbízhatóbbnak tekintett módszer a nyelvi hasonlósági (távolsági) viszonyok automatikus elemzésére és vizualizációjára. A módszer célja, hogy egy statisztikai eljárás segítségével számot adjon a nyelvi távolságok által meghatározott mintázatokról: az összetett hasonlósági viszonyrendszert kevés dimenzióban kifejezve egyszerre láttassa az összes kutatópontunk közti hasonlósági értékeket, a nyelvjárási kontinuumot. Egy kétdimenziós skálázással készült diagramon a nyelvileg nagyobb hasonlóságot mutató kutatópontok a kialakított kétdimenziós koordinátarendszerben közelebb, míg a kisebb hasonlóságot mutató kutatópontok távolabb kerülnek egymástól, ebben az esetben tehát a földrajzi távolságokat a nyelvi eltérések mértékei fölülírják (a módszerről és annak dialektometriai alkalmazásáról lásd HEERINGA 2004: 156–163; HEERINGA–NERBONNE 2013).

A 4. ábrán látható térkép háromdimenziós skálázással készült: az egyes dimenzióknak egy-egy RGB színkomponens (vörös, zöld, kék) mértéke felel meg, a kutatópontok színe így a matematikai eljárás egyes dimenzióinak mentén alakul a skálázás eredményének függvényében (lásd részletesen a GabMap honlapján: http://www.gabmap.nl/~app/doc/manual/mds.html). Ilyenkor a kutatópontok földrajzi helyzete a megszokott marad, csak az összetevőkből kikevert színek hasonlósága, illetve különbsége fejezi ki az egyes kutatópontok közti nyelvi hasonlóság mértékét. Leginkább egységes képet a partiumi, illetve a keleti székely kutatópontok mutatnak. A nyelvjárás-szigetek közül Csernakeresztúr egyértelműen a keleti székelyekkel, Lozsád mezőségi kutatópontokkal, Tatrang a tőle keletre fekvő, Székelyföld nyugati peremén elhelyezkedő kutatópontokkal, Köröstárkány Magyarkapussal, Pusztina és Diószeg a keleti székelyekkel és a két másik moldvai kutatóponttal mutat hasonlóságot. Szabóvalva nyelvi hasonlósági viszonyairól leginkább annyit tudhatunk meg, hogy eléggé különbözik valamennyi kutatóponttól. A Mezőség kirajzolódik ugyan sárgás tónusú színekkel, korántsem nyújt azonban egységesnek tűnő képet.[2]

[2] A GabMap számtalan elemzési és vizualizációs módszer használatát teszi lehetővé a felhasználó számára, amelyekkel a kutatópontok közti nyelvi hasonlósági viszonyok elemezhetők, összevethetők a földrajzi távolsággal. Jelen tanulmánynak azonban nem célja a dialektometriában alkalmazott matematikai módszerek bemutatása, erről lásd bővebben a GabMap honlapját: http://www.gabmap.nl/~app/doc/manual/references.html

4. ábra
Multidimenziós skálázással készült dialektometriai térkép a RMNyA. 1516 térképlapjának felhasználásával készített nyelvi hasonlósági mártix alapján (a térkép a GabMap szoftverrel készült, www.gabmap.nl)



4. **Konklúzió.** A teljes magyar nyelvterület dialektometriai elemzéséhez elengedhetetlen két nagyatlaszunk, a MNyA. és a RMNyA. integrálása. Az adattárak egyesítése közös adatbázisban már megtörtént, az elemzés 482 közös térképlap alapján elvégezhető. További megoldandó kérdés azonban a két adattár lejegyzési gyakorlatában mutatkozó különbségek kiküszöbölése. Enélkül a hasonlósági mátrixon végzett, a nyelvjárásterületek elhatárolását szolgáló további dialektometriai elemzésekben a lejegyzési gyakorlatban rejlő eltérések területi különbségekként jelenhetnek meg.

A dialektometriában egyre több matematikai eljárást alkalmaznak a nyelvjárások közti nyelvi hasonlósági viszonyok automatikus elemzésére. A különböző eljárásokkal készült elemzések hasonló, de nem pontosan egyező eredményei azonban nem hogy feleslegessé tennék a klasszikus nyelvészeti tudást, hanem éppen felértékelik azt. Egyre inkább fölmerül az igény, hogy föltárjuk az aggregált adatok mögött lévő valódi nyelvi struktúrákat, ötvözve az évszázados kutatói hagyományt a számítógépes nyelvészet legújabb eredményeivel.

Irodalom

GOEBL, HANS 2002. Analyse dialectométrique des structure de profondeur de l'ALF. *Revue de Linguistique Romane* 66: 5–63.

GOEBL, HANS 2005. La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de Linguistique Romane* 69: 321–367.

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.
HEERINGA, WILBERT – NERBONNE, JOHN 2013. Dialectometry. In: HINSKENS, FRANS – TAELDEMAN, JOHAN (eds.): *Language and Space. An International Handbook of Linguistic Variation, Volume III: Dutch*. (Series: Handbook of Linguistics and Communication Science (HSK)). Walter de Gruyter, Berlin and New York. 624–646.
JUHÁSZ DEZSŐ 2007. Merre tovább, magyar nyelvföldrajz? In: Zelliger Erzsébet (szerk.). Nyelv, területiség, társadalom. MNyTK. 228. Budapest. 33–43.
MNyA. = DEME LÁSZLÓ – IMRE SAMU SZERK. 1968–1977. *A magyar nyelvjárások atlasza*. I–VI. kötet. Akadémiai Kiadó, Budapest.
RMNyA. **=** MURÁDIN LÁSZLÓ gyűjt. – JUHÁSZ DEZSŐ szerk. 1995–2010. *A romániai magyar nyelvjárások atlasza*. I–XI. kötet. Magyar Nyelvtudományi Társaság, Budapest.
VARGHA FRUZSINA SÁRA 2010. A dialektometria alkalmazása és történeti helynevek nyelvföldrajzi vizsgálata a Székelyföldön. *Helynévtörténeti Tanulmányok* 5: 223–233.
VARGHA FRUZSINA SÁRA [megjelenés előtt]. A romániai magyar nyelvjárások atlasza informatizált térképlapjainak kvantitatív nyelvföldrajzi vizsgálata. In: JUHÁSZ DEZSŐ – KISS JENŐ szerk. *Egy elkészült és egy készülő magyar nyelvatlasz – kutatási tapasztalatok és perspektívák.* MNyTK. 241. Budapest.
VARGHA FRUZSINA SÁRA – VÉKÁS DOMOKOS 2009. *Magyar nyelvjárási adattárak vizsgálata interaktív dialektometriai térképekkel.* Előadás. Elhangzott a Magyar Nyelvtudományi Társaság felolvasóülésén, 2009. március 24-én.
VÉKÁS DOMOKOS 2007. Számítógépes dialektológia. In: GUTTMANN MIKLÓS – MOLNÁR ZOLTÁN szerk. V. Dialektológiai Szimpozion. Szombathely. 289–293.