ÉLŐ NYELV

Régi nyelvatlaszok – új módszerek
Nyelvtechnológiai eljárások a nyelvföldrajzban*

1. Bevezetés. – A magyar nyelvterületről eddig több mint húsz nyelvatlasz jelent meg. A nyelvföldrajzi kutatások jelenkori intenzitását jelzi, hogy a közelmúltban vehettünk kézbe több újabb regionális atlaszt is: a Nyitra-vidéki és a Sárvíz menti nyelvjárások atlaszát, valamint a Somogy–zalai nyelvatlaszt (SÁNDOR 2004.; SAJTOS 2004.; KIRÁLY 2005.). Mindemellett kiadásra várnak már évtizedekkel ezelőtt elkészült atlaszok, például TEMESI MIHÁLY munkája, az ormánsági nyelvjárások atlasza, de több, a kolozsvári iskolában készült munka is (felsorolásukat l. CS. NAGY 2004.). Hamarosan befejeződik második legnagyobb atlaszunk, „A romániai magyar nyelvjárások atlasza" (RMNyA.) kiadása.

Vannak azonban olyan atlaszaink, amelyeknek az anyaga részben már megjelent, de teljes kiadásuk még várat magára. Ezek közé tartozik a székely nyelvjárások atlasza, amelynek egy részét a „Székely nyelvföldrajzi szótár" (SzNySz.) adja közre, az atlasz nagyobb része azonban még kiadatlan (CS. NAGY 2004.). Talán kevésbé ismert, hogy „A moldvai csángó nyelvjárás atlaszá"-nak (MCsNyA.) anyaga is hasonló: ugyanis az atlasz már megjelent két kötete mellett van egy harmadik kötet is, amelynek a kiadásáról az első két kötet megjelentetésekor azért kellett lemondani, mert a kézirat elkallódott, és csak a közelmúltban került elő. A kötet informatizálására a Geolingvisztikai Műhely vállalkozott (a kutatócsoport tevékenységéről l. a Műhely honlapját: http://geolingua.elte.hu), a folyamatról a későbbiekben részletesen szólunk.

KISS JENŐ a magyar dialektológiai kutatásokat néhány évvel ezelőtt áttekintve így fogalmazott: „[i]gencsak időszerű problémakör **a gyűjtés – közzététel – feldolgozás** dialektológiabeli viszonya. Meggyőződésem, hogy a dialektológiát érintő legújabb kori kritikák nagy része [...] arra vezethető vissza, hogy a nagy és a kis nyelvjárási korpuszok (atlaszok, szótárak, szövegfelvételek stb.) kínálta lehetőségek kiaknázása, a bennük rejlő nyelvtudományi problémák megoldása általában messze elmarad az optimálistól" (KISS 2002: 13; kiemelés az eredetiben – a szerzők). KISS az eddigiekre példaként a magyar nagyatlaszt említi meg, de ide sorolhatnánk regionális atlaszaink többségét is, amelyeknek a megjelenését követően rendre elmaradt részletes feldolgozásuk. Meggyőződésünk, hogy a közreadás miként a feldolgozás lehetőségeit is meghatározza.

* A tanulmány az NKFP 5/056/2004. számú pályázata támogatásával készült. Korábbi, Vékás Domokos közreműködésével kialakított változata előadásként elhangzott a Magyar Nyelvtudományi Társaság 2005. október 25-i felolvasóülésén. Köszönetet mondunk Juhász Dezsőnek, Kiss Jenőnek és Vékás Domokosnak a dolgozat korábbi változatához fűzött megjegyzéseikért.

A következőkben amellett érvelünk, hogy alapértelmezett dialektológiai adatnak napjainkban már a számítógépen (megfelelő módon) rögzített adatot érdemes tekintenünk (BALOGH–KISS 1992.; VÉKÁS 1999.). A MCsNyA. és néhány további nyelvatlasz példáján mutatjuk be, hogyan jelentkezik minőségi különbség az informatizált adatok feldolgozásában a hagyományos információhordozókon (papíron) kiadott nyelvföldrajzi adatok felhasználási lehetőségeihez képest. Érvelésünk illeszkedik a nemzetközi dialektológia azon vonulatához, amely az informatizálási tevékenységet központi jelentőségűnek tekinti a nyelvhasználat területi vonatkozásainak feltárásában (CHAMBERS–TRUDGILL 2004: 20). A Georgiai Egyetemen például az 1980-as években kezdődött meg a múlt században gyűjtött dialektológiai adatok informatizálása az American Linguistic Atlas Project (ALAP) keretében. A számítógépes adattárak az eredeti kérdőfüzetek anyagát és magnetofonfelvételeken rögzített nyelvi adatok átírásait tartalmazzák az adatközlőkhöz kötötten. A Linguistic Atlas of Middle and South Atlantic States így rögzített és adatbázisba rendezett adataiból a Groningeni Egyetemmel együttműködésben készítettek dialektometriai elemzéseket, térképes kimutatásokat. Az adattárak egy része és a dialektometriai elemzések elérhetők a http://us.english.uga.edu/ és a http://www.let.rug.nl/~kleiweg/lamsas/ honlapokon (KRETZSCHMAR 2001.; NERBONNE–KRETZSCHMAR 2003.). A papíratlaszok anyagának számítógépes rögzítését mások is a korszerű adatelemzés elengedhetetlen feltételének tartják (PUTNAM–JOHNSON 2006.). Az újabban készülő nyelvjárási adattárak munkálatai során pedig már alapvető követelmény a megfelelő számítógépes adatrögzítés, lehetővé téve az adatok azonnali és folyamatos ellenőrzését, elemzését, feldolgozását (vö. NERBONNE–HEERINGA 2001., GOOSKENS–HEERINGA 2004.).

E tanulmányban elsősorban a MCsNyA. informatizálásának folyamatát és a munkálatok során felmerült kérdéseket, problémákat ismertetjük. A következőkben először arról lesz szó, mit értünk nyelvjárási adatok (adattárak) informatizálásán, majd ezután kerítünk sort a MCsNyA. számítógépes változata kialakításának bemutatására. A folyamat rövid áttekintése megvilágítja, hogy milyen előnyei lehetnek annak, ha régi vagy újabban készülő nyelvatlaszainkat és a számítógépes dialektológia módszereit összekapcsoljuk egymással. Végezetül a MCsNyA. informatizált változatához hasonló számítógépes adatbázisok (nyelvatlaszok és szótárak) nyelvészeti felhasználásának lehetőségeit, az adatbázisok egyesítésének szempontjait tekintjük át.

2. Nyelvjárási adatok informatizálása. – A nyelvjárási adatok informatizált formában történő rögzítésének előfeltétele a nyelvészeti szempontok által meghatározott hangjelölési kódrendszer. A kódrendszer kialakításakor elsősorban nem tipográfiai vagy esztétikai szempontokat kell figyelembe vennünk, de önmagában az informatikai szempontok érvényesítése sem elegendő. A kódrendszernek alkalmasnak kell lennie arra, hogy adatainkon hatékonyan el lehessen végezni a tudományág által meghatározott feladatok megoldásához szükséges informatikai műveleteket (statisztikai elemzések, csoportosítások, térképre vetítés, sorba rendezés stb.).

BALOGH LAJOS és KISS GÁBOR például olyan kódrendszert definiált és használt „A magyar nyelvjárások atlasza” első három kötetének informatizálásához, amelyben a szokásos betűkészletekből hiányzó alapjeleket és mellékjeleket számokkal és egyéb szimbólumokkal jelölték. Az alapjel után kitett 3-as például a félhosszúságot jelenti, a diftongusok kapcsos zárójelbe kerülnek, a diftonguson belüli nyomatékeloszlást szintén számok jelzik (BALOGH–KISS 1992., 2002.).

Segítheti az adatrögzítést, ha az informatizált anyag külalakja hasonlít a forrásdokumentuméra, vagyis a magyar egyezményes hangjelölésben használt alapjelek és mellékjelek grafikusan megjelennek. Ehhez olyan betűkészletre van szükségünk, amely külön karakterként tartalmazza a magyar egyezményes hangjelölésben használt mellékjeleket és alapjeleket, valamint olyan nyelvjárási lejegyzésszerkesztőre, azaz speciális szövegszerkesztőre, amely megfelelően támogatja a nyelvjárási szöveglejegyzést.

A Geolingvisztikai Műhelyben az adatok informatizálását VÉKÁS DOMOKOS Bihalbocs nevű dialektológiai szoftverével végezzük (honlap: http://www.bihalbocs.hu). Érdemes röviden áttekinteni, melyek voltak a legfontosabb alapelvek a lejegyzéshez használt kódrendszer és a felhasználói környezet kialakításakor.

Teljes körűen támogatja a program a magyar egyezményes hangjelölést: a ritkábban előforduló alapjelek és mellékjelek rögzítése is lehetséges, tetszőleges kombinációban; egy alapjelhez akár öt mellékjel is tartozhat. Az adatbevitelt kevés, világos elv és lejegyzésszerkesztési szabály határozza meg. Az adatok fonetikai kódok sorozatából állnak. A fonetikai alapjeleknek (pl. *e, é*) vagy mellékjeleknek (pl. félhosszúság jele, palatalizáltság jele) egyaránt külön kódok felelnek meg. A mellékjelek az alapjel után teendők ki, egyenként, valamint szintén egyenként, egymástól függetlenül levehetők az alapjelről. A mellékjelek rögzített sorrendje feltétele az adatok kereshetőségének. Több mellékjel esetén azok tetszőleges sorrendben tehetők rá az alapjelre, kötött sorrendjüket a szoftver biztosítja.

Az adatrögzítés egyszerű: speciális billentyűzet-kiosztást alkalmazunk, így minden alapjel és mellékjel egyetlen billentyű leütésével beilleszthető. Az adatbevitel azért is gyors és pontos, mivel a számítógépen rögzített anyag külalakja a hagyományoshoz hasonlít. Fontos leszögezni, hogy a megjelenés különbözősége ellenére a BALOGH LAJOS és KISS GÁBOR által használt kódolás, valamint az általunk alkalmazott, a magyar egyezményes jelek megjelenítését is lehetővé tevő kódrendszer azonos értékű, mivel informatikai módszerekkel pontosan ugyanolyan felhasználást tesznek lehetővé.

A Bihalbocsban alkalmazott kódrendszer nemcsak az adatrögzítést, hanem a feldolgozást is hatékonyabbá teszi. Megkülönböztetünk szerkesztésre, illetve keresésre optimalizált formátumokat. Kereséshez vagy csoportosításhoz például automatikusan átalakítja a program a kódokat elvontabb vagy fonetikusan „durvább” formába. Szerkesztésnél az *sz* hangot az s és z billentyű leütésével rögzítjük, mert így a legegyszerűbb; keresésnél azonban más a helyzet: ha az *s* hangot keressük, vagy ha az *s* mássalhangzók számára vagyunk kíváncsiak, akkor az *sz* hangokban grafikusan meglévő *s*-eket nem akarjuk megtalálni. Továbbá, ha a rövid *s* hangokra vagyunk kíváncsiak, akkor a hosszú *s* hangokat szintén nem akarjuk figyelembe venni. Adatok csoportosításakor úgy is dönthetünk, hogy figyelmen kívül hagyjuk a mellékjeleket, és akár összevonhatjuk a hasonló magánhangzókat is, hogy kevesebb csoportosítandó adatváltozathoz jussunk (pl. az *œ* és *ö* hangokat egyaránt *ö*-nek vesszük).

A hangfájlokból történő lejegyzés során a hangrészletek megszólaltathatók és oszcillogramként (hullámformában) megjeleníthetők. A speciálisan dialektológiai szövegek tárolására kialakított BXT-fájlformátumban pedig lehetőség van a lejegyzés és a hangfájl szinkronizálására, vagyis időzítési markerek utalnak a szövegben a hangfájlok vonatkozó időpillanataira (a nyelvjárási adatok informatizálásáról bővebben l. VÉKÁS 1999.).

Az adatokat megfelelően kódolva, strukturált szövegfájlokban tároljuk. A fájlok fejléce tartalmazza mindazon információkat, amelyekkel a kiindulási térképlapról (szócikkről) rendelkezünk a forrásdokumentum alapján (címszó, térkép száma, kérdőmondat stb.). A fej-

lécet követően az adatokat kötött struktúrában soroljuk fel. Az eddigiekben kétféle struktúrát határoztunk meg, hogy jobban követhessük a forrásdokumentumok felépítését, és így hatékonyabbá tegyük az adatrögzítést. Az egyik a kutatópontok szerint szerveződik, vagyis minden kutatópontot azonosítóval látunk el, és a kutatópont azonosítója (sorszáma, betűjele) mellett rögzítjük az adott kutatóponthoz tartozó adatokat. A másik szerkezet az adatok szerint szerveződik, vagyis az adatok alapján alakítjuk ki a sorrendet, és minden adat mellett fölsoroljuk azoknak a kutatópontoknak a kódját, ahol az adat előfordul.

3. A moldvai csángó nyelvjárás atlaszának informatizálása. – Az atlasz számítógépes változatának elkészítése szoros összefüggésben áll kiadásának történetével. A MCsNyA. kéziratának történetéről újabban TÁNCZOS VILMOS készített összefoglalót (TÁNCZOS 2004: 217–9), így itt csak az informatizált változat elkészítése szempontjából fontos tényezőket említjük meg.

A MCsNyA. mindmáig a legjelentősebb vállalkozás volt a moldvai magyar nyelvjárások vizsgálatára, azóta is megkerülhetetlen forrása minden moldvai nyelvföldrajzi kutatásnak. Az atlasz terepmunkája 1949–1962 között zajlott, de legjelentősebb része már 1957-re megvalósult. A terepmunka fő szakaszát követően a kutatás vezetője, SZABÓ T. ATTILA arról számolt be, hogy ekkorra elkészült „mintegy 900 térképlap sokszorosításra alkalmas tusrajzban" (SZABÓ T. 1959: 37). Ez akkor még nem a tervezett atlasz teljes anyaga volt, az összesen 1100-1200 térképlapból állt volna (SZABÓ T. 1959: 31). Ma ezt a 900 térképlapnyi anyagot tekintjük a MCsNyA. kéziratának. A terepmunka fő szakaszát követően a kutatócsoport tagjai legközelebb 1962-ben folytatták a kutatást Moldvában, de ekkor már csak ellenőrző gyűjtéseket végeztek.

Ami az informatizált változatában most teljessé vált atlaszból kimaradt, jelenleg a kolozsvári Babeş–Bolyai Tudományegyetem Magyar Nyelvészeti Tanszékén található meg kicédulázva, de más moldvai kutatások szóanyagával együtt – a készülő csángó szótár adatbázisának részeként (l. PÉNTEK 2004.). Ennek az anyagnak a nyelvföldrajzi hasznosítására jelen kutatásunkban nem vállalkozhattunk.

A már megrajzolt és kiadásra előkészített 900 térképlap megjelentetésére hosszú időn át nem adódott lehetőség, és csak az 1990-es évek fordulóján kezdődhetett meg a Budapestre eljuttatott térképlapok közreadása. A MCsNyA. I. és II. kötete 1991-ben jelent meg Budapesten. Sajnálatos módon azonban a kiadás folyamatába hiba csúszott: a nyomdában olyan térképlapokat hoztak létre, amelyek az adatok mellett nem tüntették fel a kutatópontok számát. Mivel a megjelent kötetekben hiányzott a közreadott adatok mellett a lokalizálásukat elősegítő számozás, a főként a Szeret mentén sűrűn egymás mellé települt moldvai falvak adatbeírásos módszerrel megjelenített nyelvjárási adatait csak nehezen lehetett pontosan egy-egy kutatóponthoz kötni.

A már megjelent kötetek informatizálásának az volt az egyik speciális indoka, hogy lehetővé váljon a kutatópontok és az adatok pontos egymáshoz rendelése, és ennélfogva az atlasz felhasználását ne nehezítsék meg a papíron kiadott változat hiányosságai.

Ezen a speciális szemponton kívül további érvek is szükségessé tették a számítógépes változat létrehozását. Ezeket a MCsNyA. III. kötetének informatizálása során felvetődött feladatokkal kapcsolatban mutatjuk be. Mindehhez azonban ismét szükséges kitérni az atlasz keletkezésének körülményeire.

A III. kötet sorsáról Murádin László visszaemlékezése alapján TÁNCZOS VILMOS a következőket írta: „a kiadás időpontjában (1991) a 3. kötet kéziratát a szerkesztők sehol nem találták. A kézirat az 1987-ben elhunyt Szabó T. Attila könyvtárában egy szekrény mögé volt elrejtve, és csak évekkel később került elő, akkor, amikor a nyelvészprofesszor örököse a könyvtár anyagát az Erdélyi Múzeum-Egyesületnek adományozta, és a könyvtár elszállítására is sor került" (TÁNCZOS 2004: 264).

Az így előkerült III. kötetet 2005 tavaszán vettük át Péntek Jánostól Kolozsváron, és hamarosan megkezdtük a már számítógépre vitt első két kötetből álló adatbázis kibővítését az újabb adatokkal. A III. kötet felépítése azonban több szempontból is eltért az előbbi kötetekétől. A különbségek közül a leginkább szembetűnő az volt, hogy a kiadatlan térképlapok jelentős része ún. „összesítő térképlap" – ma úgy mondanánk: jelenségtérkép – volt, amely több térkép adatait valamely szempont alapján egyetlen lapra sűrítve mutatta be. Ezt szemlélteti a következő mutatvány, amely a III. kötet első öt térképlapjának címszavait sorolja fel (a kötet teljes címszólistája BODÓ CSANÁD és ERIS ELVIRA összeállításában megtekinthető a http://geolingua.elte.hu/projects/mcsnya3cimszok.pdf internetes címen, l. még BODÓ–VARGHA 2007.):

609. ***emberem* (609)**; *fecske* (520); ***vejem* (610**) – az *e* hang ejtése hangsúlyos helyzetben;
610. ***kilencven* (611)**; *este* (566); *egyenes* (297) – az *e* ejtése hangsúlytalan helyzetben;
611. *ég* <ige> (33); *ég* <főnév> (461); *éjfélkor* (569) – az *é* ejtése hangsúlyos helyzetben;
612. *fehér* (272); *gané* (317: trágya); ***ehetném* (612)** – az *é* hangsúlytalan helyzetben;
613. *pepeny* 'sárgadinnye' (356: sárgadinnye); *veverica* 'mókus' (545: mókus); ***berbécs*** 'kos' **(613)** – az *e*-féle hangok ejtése román kölcsönszavakban.

A III. kötet kézirata a 609. térképlappal kezdődik. Ez rögtön három adatsort összesít (az *emberem*, a *fecske* és a *vejem* címszó moldvai megfelelőit), és jelenségtérképként a köznyelvi *e* nyelvjárási megfeleléseit mutatja be hangsúlyos helyzetben. Az utána következő térképlapok is három-három térkép összesített adatait adják közre. A kiemelés nélkül álló címszók nem újak, már jelen voltak az atlasz I. vagy II. kötetében; a félkövér betűvel szedett címszók ellenben a megjelent kötetekhez képest új információkat jelenítenek meg. A kiemelt címszók után zárójelben álló számok az atlasz informatizált változatához kialakított megfelelő adatfájlok sorszámai. Az első két kötetben ezek a számok a hagyományos és az informatizált változat térképlapjain azonosak egymással. (A 613. kéziratos térképlap rámutat arra, hogy nem feltétlenül a címszók és a térképlapok azonosságáról van szó, ugyanis pl. a köznyelvi *mókus*-nak az atlasz második kötetében már közreadott nyelvjárási megfelelői közül a jelenségtérkép csak azokat ábrázolja, amelyek a román *veveriță* átvételei. Ezekben az esetekben az első két kötet megfelelő térképlapjának számát követően megadjuk az ott feltüntetett címszót is.)

A harmadik kötet kéziratos és számítógépes változatának térképlapjai már eltérnek egymástól. Ez abból a döntésünkből adódik, hogy nem követtük a kézirat szerkezetét, és a jelenségtérképek adatait nem egyetlen fájlban rögzítettük, hanem külön-külön adatsorokat alakítottunk ki. Például a kézirat 609. térképlapjából két új fájlt hoztunk létre, az egyiket a 609. sorszámot viselő *emberem* címszóval, a másikat 610. számmal *vejem* címszóval. A *fecske* adatainak rögzítésére ekkor már nem volt szükség, mivel azokat korábban számítógépre vittük a II. kötet informatizálásának részeként.

Ha összevetjük a már megjelent kötetek anyagával, akkor a III. kötetben az előző kötetek anyagát jócskán meghaladó mennyiségű új térképlap található: a MCsNyA. I. kötetében 295, a II.-ban 303 térképlap található, míg a III.-ban 437 új térképlap van. (A legutóbbi kötet papíron létrehozott kéziratos változatában a térképlapok száma ennél jóval kisebb, mivel számos új adat összesített térképlapokon jelenik meg.) Az atlasz teljes informatizált adatbázisa táblázatos formában hozzáférhető a MCsNyA. anyagából készített jelenségtérképeket bemutató CD-kiadványunk mellékleteként (BODÓ–VARGHA 2007.).

Miért döntöttünk úgy, hogy a szerkesztők eredeti szándékától eltérve az adatrögzítés nem jelenségtérképek szerint valósul meg? Gyakorlati megfontolásból például a mutatványban szereplő 611. térképlap adatainak ismételt rögzítése felesleges, mivel azok már szerepelnek az atlasz korábbi térképlapjain. Ugyanakkor az általuk bemutatott nyelvjárási jelenséget – itt éppen az *é* dialektális megfelelőit hangsúlyos helyzetben – egy informatizált nyelvatlaszban gyorsan térképre vetíthetjük, sőt, más jelenségeket bemutató térképeket is készíthetünk akár ugyanezeknek az adatoknak a felhasználásával. (Az atlaszkötet feldolgozása során adódott további kérdésekről l. BODÓ–VARGHA 2007.)

Ezeket a lehetőségeket a kéziratos atlasz 618. térképlapján mutatjuk be. A térkép három korábbinak az adatait egyesíti (zárójelben a térképlapok száma): *füst* (36.), *süt* (221.) és *süket* (259.). A kézirat szerint az összesített térkép által bemutatandó jelenség a köznyelvi *ü* nyelvjárási megfelelőinek területi megoszlása. A moldvai *ü* : *i* megoszlást a három térkép informatizált változatának mechanikus egyesítésével létrehozott jelenségtérkép mutatja be, amely a felső nyelvállású magánhangzó labialitása alapján csoportosított adatokból automatikusan generált módon jött létre.

A három térképlap integrált változata az 1. ábrán látható. A halvány szürkék az *ü*-féle hangokat tartalmazó alakok, a sötétebb színek pedig az *i*-féle magánhangzós alakok. A mellékjeles átmeneti alakokat az alapjellel jelölt fonémához soroltuk. A térképen a jelmagyarázatban megadott alakváltozat-párokat szimbolizáló színek a kutatópontok körcikkeinek az óramutató járásával megegyező sorrendjében jelennek meg. (Az itt bemutatott térképek eredeti, színes változatát elhelyeztük a Geolingvisztikai Műhely honlapján. Az informatizált térképeket böngészve nemcsak a kutatópontok nevét, hanem a nyelvi adatokat is leolvashatjuk, ha a kurzort egy-egy szimbólum fölé húzzuk. Az 1. ábra színes, internetes változatának lelőhelye: http://geolingua.elte.hu/maps/mcsnya/fustfist/terk_frames.html.) Moldvai térképeink adatainak lokalizálását elősegítendő a 2. ábrán mutatjuk meg a MCsNyA. kutatópontjainak sorszámát, majd táblázatban soroljuk fel a települések nevét. (Névhasználatunkban nem követjük minden esetben az MCsNyA. I. kötetében feltüntetett településneveket, ennek indokairól l. BODÓ–VARGHA 2007.)

A három térképlap egyesítésével létrehozott integrált térkép elemzési lehetőségei a köznyelvi *ü* nyelvjárási megfelelőinek vizsgálatával nem merülnek ki, például felvethető egy további hangtani kérdés is: hol lép fel a köznyelvi *s* helyén a moldvai csángó nyelvjárások egyik sztereotipikus jegye, az *sz*-ezés. A 3. ábra e szempont szerint csoportosítva mutatja be a három térképlap adatait (színes, internetes változat: http://geolingua.elte.hu/maps/mcsnya/fustfuszt/terk_frames.html). Itt a halványszürke színek a köznyelvinek megfelelő *s*-ek előfordulását jelölik, az ezeknél sötétebb színek az *sz*-féle hangokat.

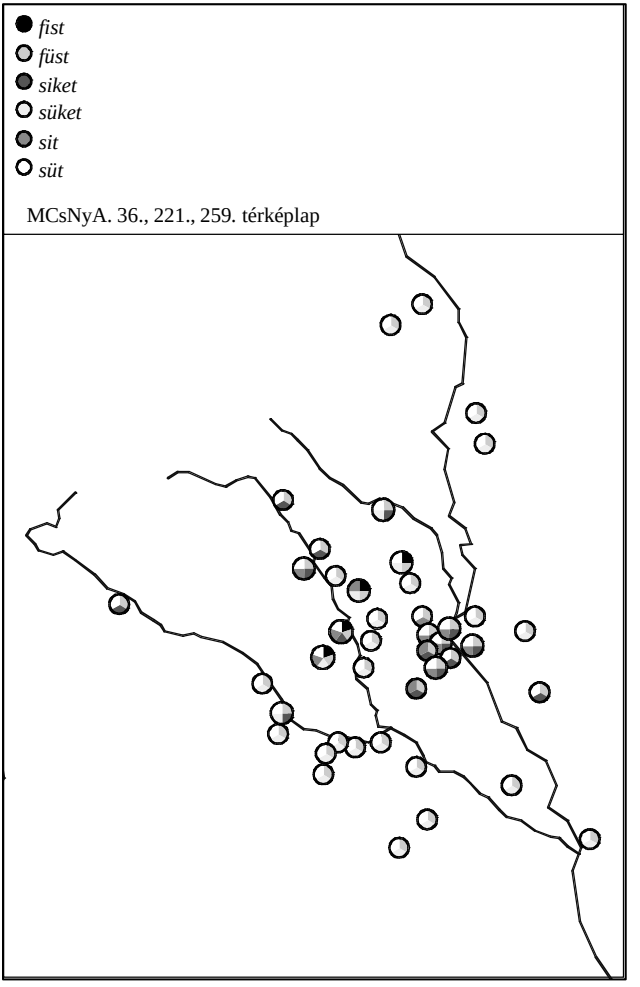
Az összesített jelenségtérképek alapján elmondhatjuk, hogy az illabiális ejtésű adatok leginkább a Szeret völgyében és néhány Tázló menti településen fordulnak elő, míg egyáltalán nincsenek jelen az északi csángó településeken, és a Tatros menti falvakban sem jel-

lemzők. Az *sz*-ezés területi megoszlása várakozásainknak megfelelően alakul: a jelenség az északi és a déli csángó nyelvjárásokban van jelen, de egyetlen *sz*-ező adatot találunk Balanyászán is. Megjegyzendő, hogy míg az északi csángó településeken szinte kizárólag *sz*-ező alakokat találunk, addig a Szeret menti településeken általában egymás mellett jelenik meg az *s*-es és az *sz*-ező változat az adatokban.

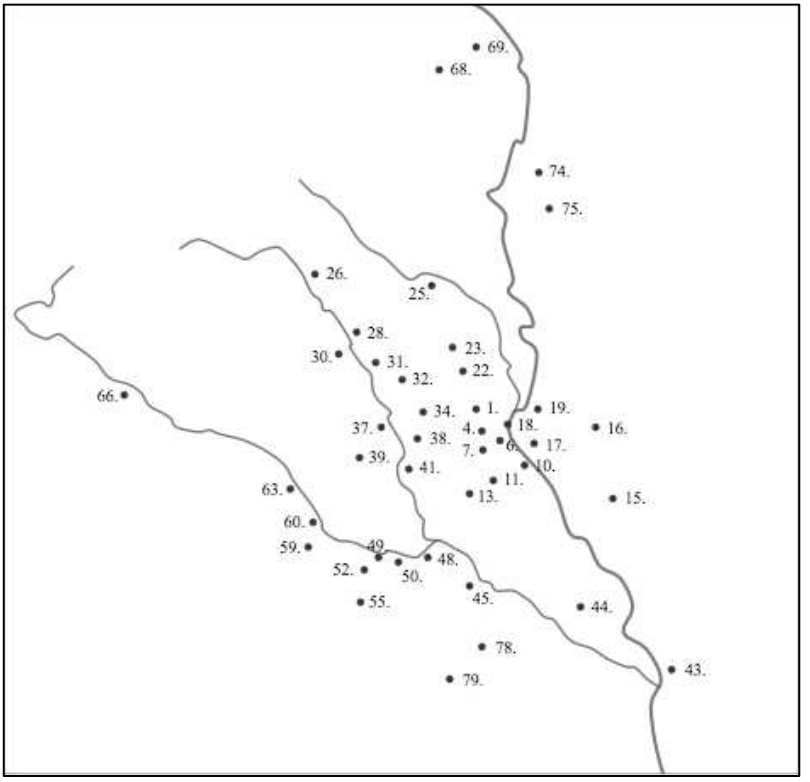
A MCsNyA. harmadik kötetét tehát úgy informatizáltuk, hogy az esetleges papírkiadású változat létrehozását későbbre hagytuk. Ezt a döntésünket – amellett, hogy az első kettő és a harmadik kötet között igen sok az átfedés – az informatizált atlaszok egyesítésének lehetősége is indokolttá teszi. A következőkben a nyelvatlaszok és más adatbázisok integrálásának jelentőségét világítjuk meg néhány példával.

1. ábra

Az *ü* ~ *i* megoszlás Moldvában a MCsNyA. *füst* (36.), *süt* (221.) és *süket* (259.) térképlapján



2. ábra
A MCsNyA. kutatópont-hálózata

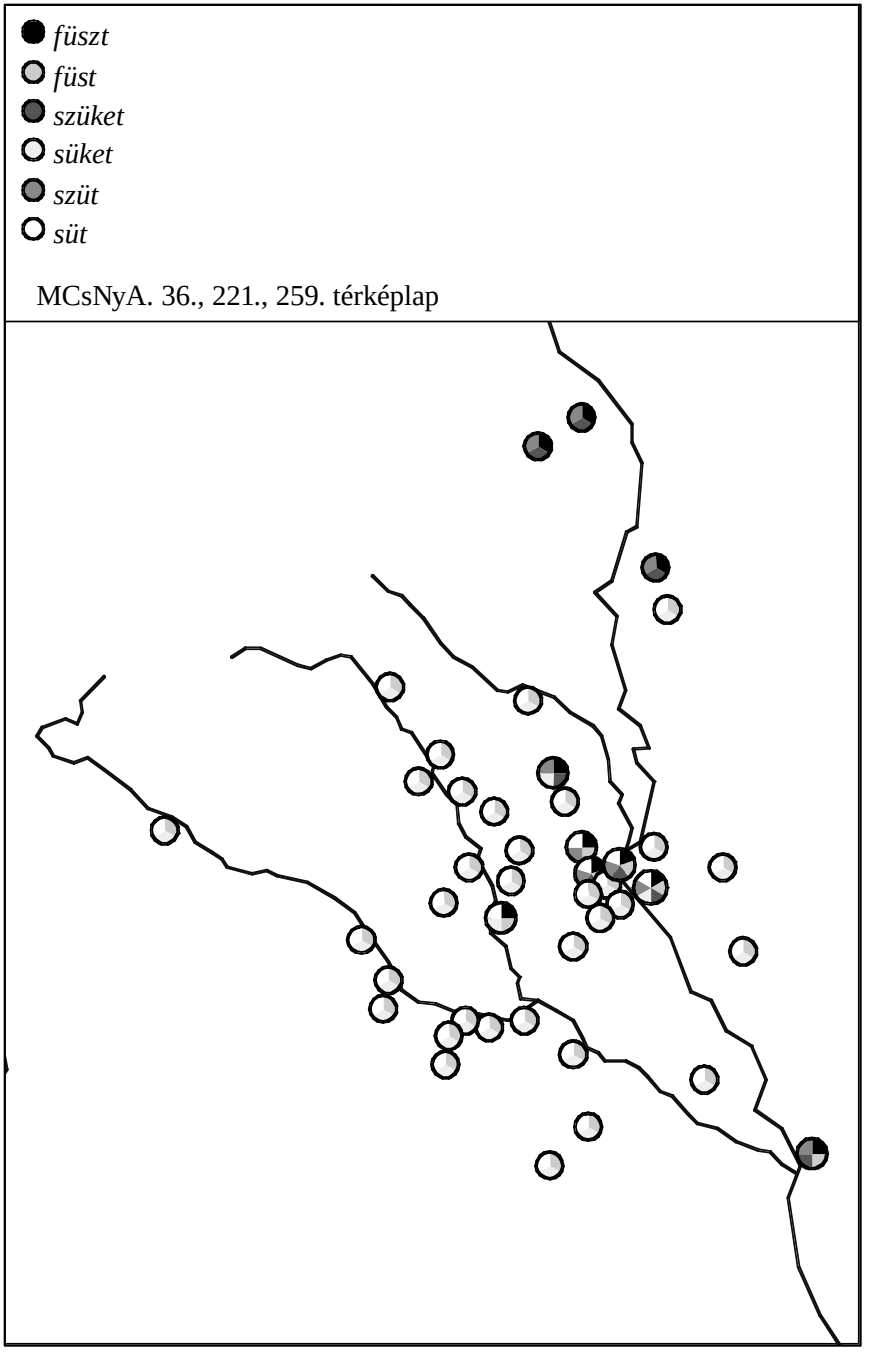


A MCsNyA. kutatópontjainak jegyzéke:

1. Bogdánfalva
4. Nagypatak
6. Kákova (Forrófalva)
7. Klézse
10. Dózsa
11. Csík
13. Külsőrekecsin
15. Magyarfalu
16. Lábnik
17. Gyoszény
18. Trunk
19. Ketris
22. Lujzikalagor
23. Szakatura
25. Lészped
26. Frumósza
28. Pusztina
30. Szoloncka
31. Szerbek
32. Esztrugár
34. Gajdár
37. Gyidráska
38. Esztufuj
39. Berzunc ~ Butukár
41. Balanyásza
43. Ploszkucén
44. Szászkút
45. Válészáka
48. Onyest
49. Tatros
50. Diószeg
52. Újfalu
55. Gorzafalva
59. Szalánc (Templomfalva)
60. Pakura
63. Dormánfalva
66. Csügés
68. Kelgyest
69. Szabófalva
74. Balusest
75. Kalugarény
78. Prálea
79. Vizánta

3. ábra

Az *s* ~ *sz* megoszlás Moldvában a MCsNyA. *füst* (36.), *süt* (221.) és *süket* (259.) térképlapján



4. Adatelemzés, adatbázisok integrálása. – Az adatrögzítést bármilyen szövegszerkesztővel elvégezhetjük, elegendő, ha megfelelően pontos kódrendszert definiálunk. Az adatkezeléshez azonban már speciális szoftverre van szükség. Kereskedelmi forgalomban nincs olyan program, amellyel összetett geolingvisztikai feladatokat el lehetne végezni. Ebből következik, hogy a tudományág feladata a fejlődéséhez szükséges informatikai eszközök megalkotása. Az általunk használt szoftver – habár a fejlesztők célja egyelőre elsősorban a hatékony adatrögzítés lehetővé tétele – további lehetőségeket biztosít az adatok különféle megjelenítésére, feldolgozására, elemzésére is.

Az adatelemzés feltétele, hogy informatizált adatokat (adattárakat) alakítsunk ki. Kutatócsoportunk évek óta foglalkozik nyelvjárási adattárak informatizálásával. A MCsNyA. mellett rögzítettük „A magyar nyelvjárások atlasza" (MNyA.) 4–6. kötetét. BALOGH LAJOS és KISS GÁBOR egy másik kódrendszer használatával már korábban informatizálta a MNyA. első három kötetének nagy részét. Módszerük teljes értékűnek mondható, a gyakorlatban azonban az adatrögzítés során alkalmazott számos különböző platform kompatibilitási problémákat okozott. Rögzítettük továbbá a SzNySz. teljes anyagát. Érdemes megemlíteni, hogy ebben a szótárban az eredeti székelyföldi anyag jó része nem jelent meg, hanem cédulák, illetve kéziratos térképlapok formájában még informatizálásra vár (CS. NAGY 2004.). Jelenleg még tart a RMNyA. szerkesztése és kiadása JUHÁSZ DEZSŐ irányításával, ezzel párhuzamosan kezdtük meg az informatizált változat kialakítását.

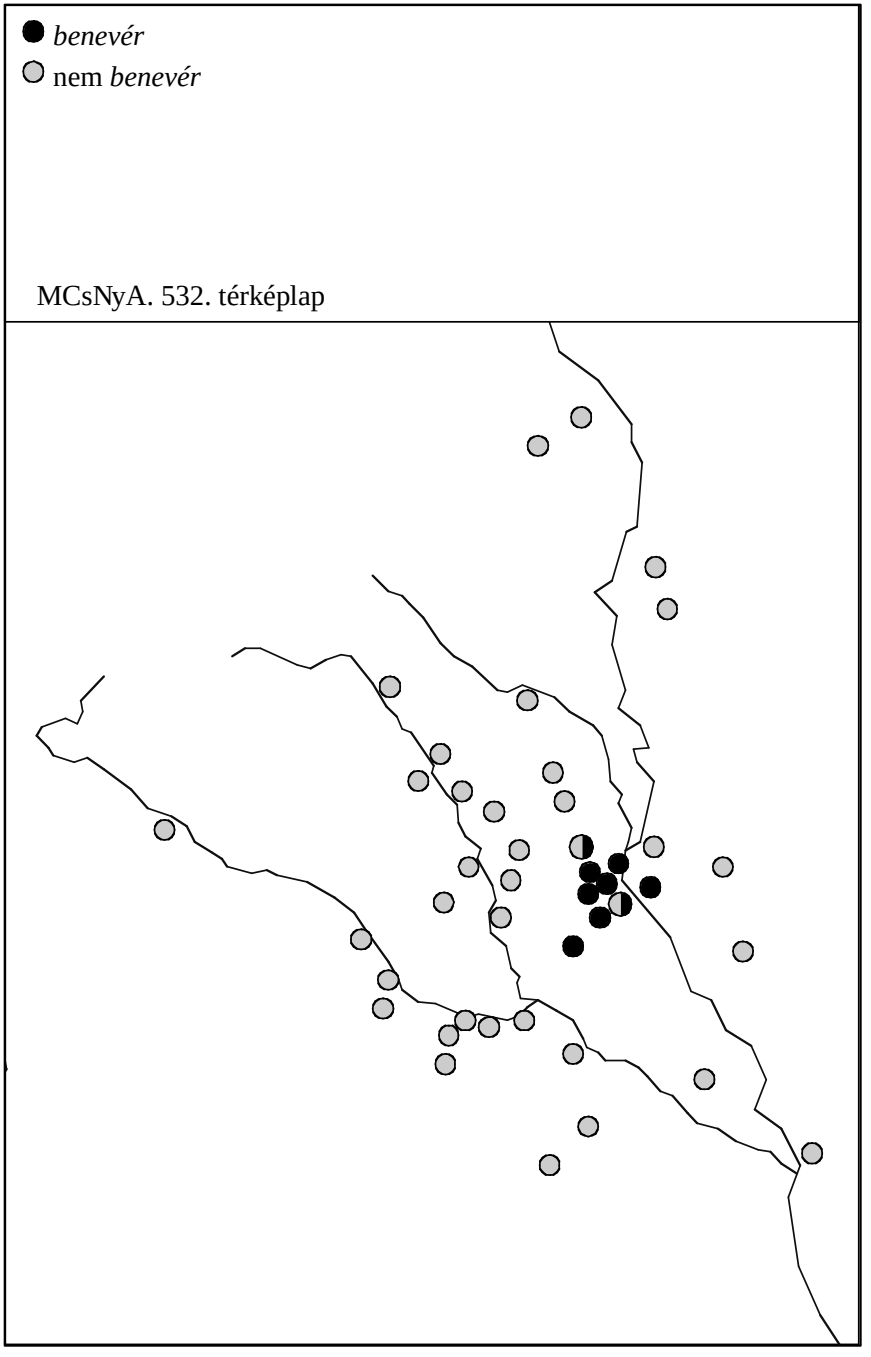
Az informatizált adatok egyik előnye, hogy integrálhatók. Integrálásnak nevezzük a különböző adattárakból, például térképlapokról származó adatok valamilyen – általunk meghatározott – szempontok szerint történő összesítését. Ha összesítjük a különböző adattárakból származó informatizált adatokat (pl. egy közös térképen), nem veszíthetik el a rájuk vonatkozó információkat (kutatóponthoz, adatközlőhöz való kötődésüket, minősítéseiket, egyéb megjegyzéseket stb.).

Az integráláshoz két feltételnek mindenképp teljesülnie kell. Integrálni egyrészt azonos vagy kompatibilis kódolású adatokat lehet. A különböző pontossági szinten lejegyzett adatok a kevésbé pontos (pl. mellékjelek nélküli) lejegyzés szintjén lesznek kompatibilisek, vagyis összevethetők. Másrészt különböző adattárak integrálására, egy közös térképen való megjelenítésére csak akkor van mód, ha egyesített kutatópont-hálózatot alakítunk ki.

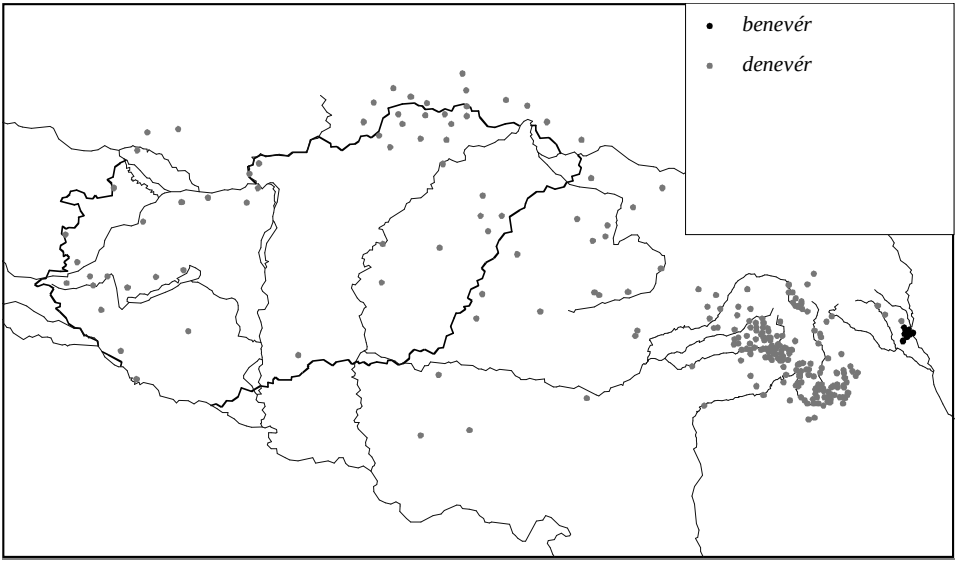
Az integrálásból adódóan lehetőségünk van a különböző adattárakból származó adatok egyetlen térképen való megjelenítésére, csoportosított térképek, hangstatisztikák készítésére. Készíthetünk például adatlistát egy-egy kutatópont adataiból akár több adatbázis alapján; megjeleníthetjük egy térképen különböző atlaszok és nyelvföldrajzi szótárak adatait; készíthetünk hangstatisztikát egy kutatópont adataiból; térképre vetíthetünk olyan kimutatásokat, amelyek egy adott hang előfordulási arányait szemléltetik az általunk meghatározott kutatópont-hálózat és adattár(ak) szerint.

A 4. ábra térképét a MCsNyA. *denevér* címszavú térképének adataiból készítettük. A térképen feketével jelöltük a *benevér* adatokat. Jól látható a *benevér* alak előfordulásának területi megoszlása Moldvában. Arról azonban, hogy a magyar nyelvterületen előfordul-e máshol is *benevér*, csak úgy győződhetünk meg, ha integráljuk a MCsNyA. adatait más adatbázisokkal. Az 5. ábrán a MCsNyA., a MNyA., a RMNyA. és a SzNySz. alapján azok a kutatópontok látszanak, amelyeken előfordul a *denevér* lexikai változat. Feketével jelöltük a *benevér*, szürkével pedig a *denevér* hangtani változatot. Az integrált térkép alapján megállapítható, hogy a *benevér* egyedül Moldvában fordul elő, ennélfogva valószínűleg belső keletkezésű változat. (A térkép készítésekor az újkeletűként minősített *denevér* adatokat nem vettük figyelembe.)

4. ábra
A *benevér* elterjedtsége Moldvában



5. ábra
A *denevér ~ benevér* megoszlása a magyar nyelvterületen



A Bihalbocs geolingvisztikai szoftver alkalmazásával nemcsak azonos vagy hasonló címszavú térképek egyesítésére van lehetőségünk, hanem különböző fogalmak megnevezését bemutató térképlapok adatainak csoportosítására is vállalkozhatunk. A 6. ábrán megjelenített térképet például a MNyA., a RMNyA. és a MCsNyA. *patkány* és *vakondok* címszavú térképlapjaiból, valamint a SzNySz. megfelelő szócikkeiből hoztuk létre. Összesen tehát hat térképlapról és a SzNySz. két szócikkéből származnak az adatok. A térképen csak azok a kutatópontok látszanak, ahol használják a *patkány* lexémát valamilyen jelentésben. A világos színnel megjelölt kutatópontokon a *patkány* a köznyelvi jelentésben él, a sötéttel megjelölt kutatópontokon azonban 'vakondok' jelentésű.

A különböző adattárak integrálásával lehetővé válik a jelenségnek a magyar nyelvterület egészén való szemléltetése. A 'vakondok' jelentésű *patkány* földrajzi megoszlásának tágabb összefüggéseit 6. ábránk olyan módon szemlélteti, hogy az eddig négy különböző kiadványban több térképlapon bemutatott adatokat egyetlen térképen jeleníti meg. A SzNySz. adatait a jobb áttekinthetőség érdekében külön térképlapon is szemléltetjük (7. ábra).

A SzNySz.-ból tájékozódó olvasó a *patkány* 'vakondok' jelentését a nyugati székelységre jellemző nyelvjárási sajátosságként azonosíthatja. A RMNyA.-t kézbe véve a nem köznyelvi jelentésű lexéma területi elterjedéséről már azt is megtudhatja, hogy a keleti nyelvjárási régiók közül csak a Mezőségen általános a használata. Ebből le lehetne vonni azt a következtetést, hogy a *patkány* 'vakondok' jelentése a nyugati székely nyelvjárásokban mezőségi hatásra terjedt el (HEGEDŰS 2002: 123). Ám azt is feltehetjük, hogy a jelentés éppen a mezőségi nyelvjárásban számít keletről átvett újításnak, mivel térképünk tanúsága szerint a feltételezett újítás nincsen jelen az archaikusabb nyelvjárássziget-helyzetben lévő mezőségi kutatópontokon, mint pl. Lozsád, Halmágy, Domokos vagy Köröstárkány. Ebből

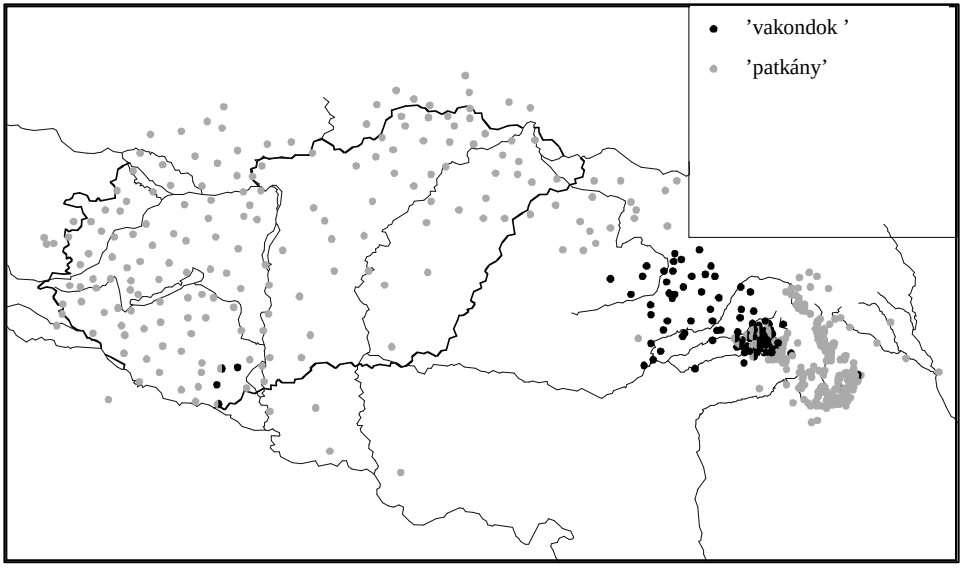
a szempontból a jelenség moldvai hiánya nem perdöntő: ha BENKŐ (1989.), majd JUHÁSZ (2004.) nyomán a csángó nyelvjárások mezőségi gyökereit kívánjuk kimutatni, a valószínűsíthető történeti kapcsolatokat a vizsgált moldvai falvakban elfedi a román nyelvi hatásra kölcsönszóként megjelent *guzgán* 'vakondok' lexéma általános használata. (Erre utal a MCsNyA. északi és déli csángó kutatópontjainak hiánya a térképen.)

A *patkány* 'vakondok' jelentésének elsődleges mezőségi kialakulását egy további szempont miatt sem tarthatjuk valószínűnek. A jelentés nyugati székely eredete mellett szóló érv, hogy a MNyA. térképlapjain a jelentés dél-dunántúli gócpontja is megjelenik az eddigiek mellett. A jelentés a nagyatlasz négy kutatópontján (Martonfa, Szebény, Kisherend és Old) fordul elő (az ÚMTsz. szerint adatolták a közeli Kopácson is). Ez a területi megfelelés egybevág BENKŐ (1967.) nyelvföldrajzi megalapozású feltevésével a Dráva alsó folyásának vidékén beszélt nyelvjárások és az udvarhelyi székely nyelvjárás történeti kapcsolatáról, és egyben további érvet jelent amellett, hogy a jelenség nem a mezőségi nyelvjárásból került át a nyugati székelység nyelvébe.

Az informatizált adatok felhasználásának további lehetősége, hogy a meglévő adatokból hangstatisztikai térképet készíthetünk. A 8. ábra például nyitódó *(i͡ë, ü͡ö, u͡o)* és záródó *(ë͡i, ö͡ü, o͡u)* diftongusok előfordulási arányait mutatja Moldvában és a Székelyföldön, a SzNySz. és MCsNyA. adatai alapján. (A térképen fekete színnel jelöltük a nyitódó, szürkével pedig a záródó diftongusokat.)

6. ábra

A *patkány* jelentései a magyar nyelvterületen



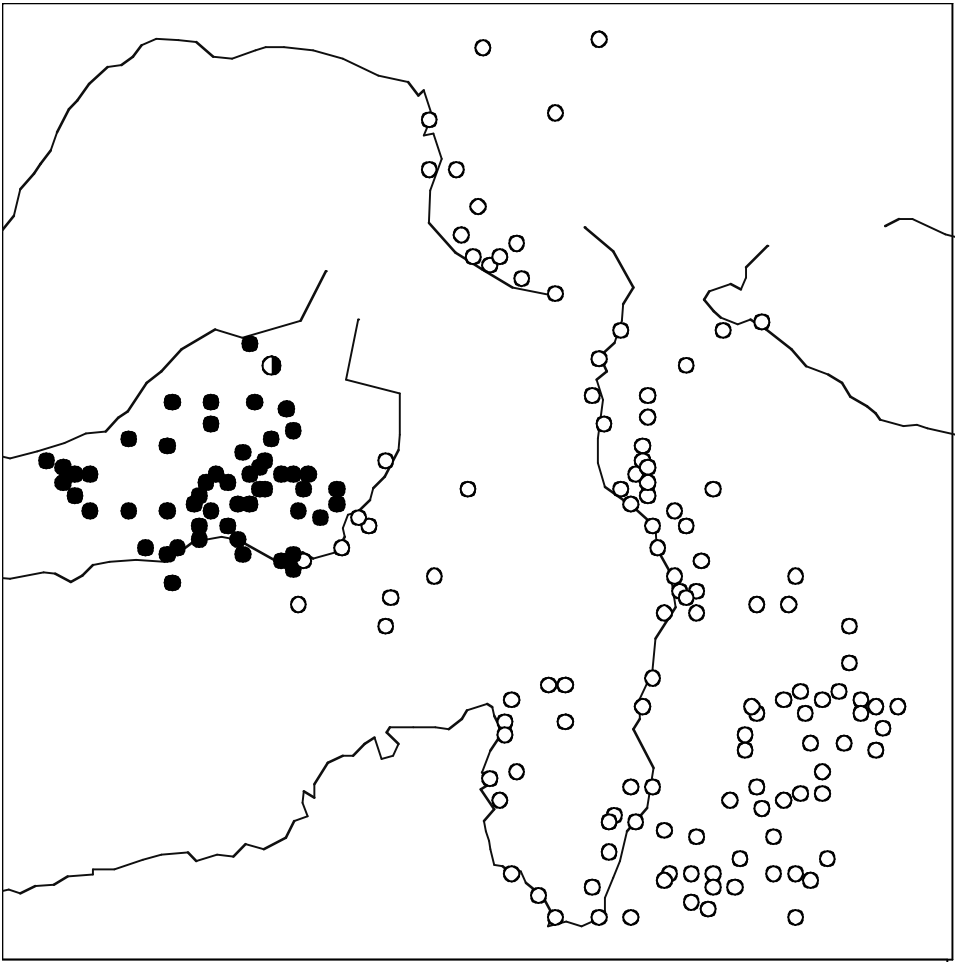
Azokon a kutatópontokon találunk teljes (100%-ban fekete vagy szürke) kört, ahol a vizsgált adatokban az adatok számához viszonyítva legtöbbször fordulnak elő diftongusok (záródó diftongusok Zabolán vannak jelen a legnagyobb arányban, ezért itt teljesen fekete

a kör). A többi kutatóponton ezen kutatópontokhoz viszonyítva adjuk meg a diftongusok arányát. (A térkép generálásánál nem voltunk tekintettel a diftongus elemei közötti nyomatékmegoszlásra.) Az ilyen jellegű vizsgálatok lehetőséget adhatnak a különböző kutatópontok hangrendszerei között meglévő összefüggések feltárására és vizsgálatára.

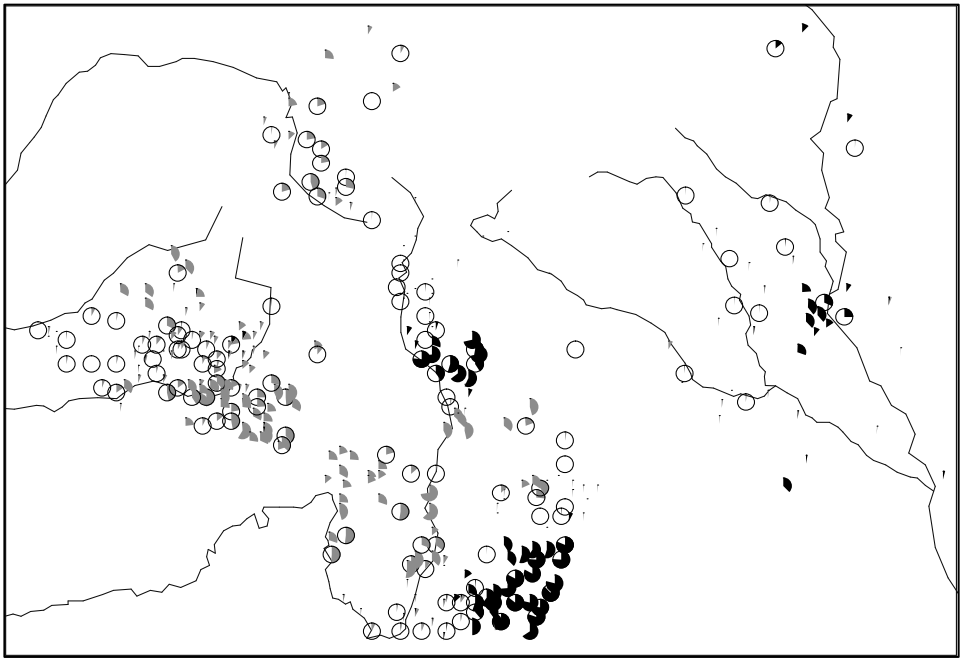
Az informatizált nyelvatlaszok számos lehetőséget kínálnak a kutatásra az itt bemutatottakon túl is. A RMNyA. adatainak fonetikai szempontú elemzésére l. VÉKÁS 2000., a MCsNyA. anyagán végzett hangstatisztikai elemzésre l. BODÓ 2007., a Szilágysági nyelvatlasz (MÁRTON–HEGEDŰS 2000.) hangtani jelenségeinek vizsgálatára l. HEGEDŰS 2007. Néhány további példát láthatunk a Geolingvisztikai Műhely említett honlapján.

7. ábra

A *patkány* jelentése a SzNySz.-ban (’vakondok’ = fekete, ’patkány’ = fehér)



8. ábra
Nyitódó (fekete) és záródó (szürke) diftongusok előfordulási aránya
a magyar nyelvterület keleti régióiban



5. Összefoglalás és kitekintés. – A fentiekben a MCsNyA. Informatizálásának folyamatával szemléltettük, hogy milyen lehetőségek rejlenek hagyományos adattáraink új módszerekkel való feldolgozásában. Egy-egy nyelvjárási adattár informatizálásával az adott atlasz vagy szótár adatainak értéke többféle okból is jelentősen megnő. Amellett, hogy az anyag alkalmassá válik az informatikai módszerekkel történő feldolgozásra (statisztikai kimutatások, adatok csoportosítása és térképre vetítése, különböző térképlapok egyesítése, a világhálón megjeleníthető térképek kialakítása stb.), az adatokat integrálni lehet más adattárakban szereplő adatokkal is.

Dolgozatunkban néhány példával mutattuk be, hogy az informatizált adatok és adattárak már most is számos érdekes összevetést, elemzést tesznek lehetővé. Adatbázisunk azonban korántsem teljes. Két legjelentősebb nyelvföldrajzi atlaszunk, a MNyA. és a RMNyA. teljes anyagát sem lehet még informatikai eszközökkel kutatni, és természetesen nem szabad megfeledkeznünk azokról a kisebb, de jelentős nyelvföldrajzi atlaszokról és szótárakról sem, amelyeknek még meg sem kezdődött a számítógépes feldolgozása. Egyik legfontosabb feladatunk, hogy ezt az elmaradást minél előbb felszámoljuk.

A már meglévő adattárak feldolgozása mellett azonban tovább kell folytatni az adatgyűjtést, kihasználva azokat az új technikai megoldásokat, amelyek lehetővé teszik hanganyag és lejegyzés szinkronizálását (összekapcsolását), valamint hatékonyan segítik az adatok adatbázisba rendezését. A jövőben első lépésként olyan informatizált adattárakat

célszerű kialakítani, amelyekből a következő munkafázisokban már viszonylag egyszerűen létrehozhatjuk az adatok egyéb megjelenítési formáit (nyomtatott térképlapok, internetes anyagok, hanglemezek stb.) a különböző szintű és jellegű igényeknek megfelelően.

A hivatkozott irodalom

BALOGH LAJOS – KISS GÁBOR 1992. A magyar nyelvjárások atlaszának számítógépes feldolgozása. In: KONTRA MIKLÓS szerk., Társadalmi és területi változatok a magyar nyelvben. MTA Nyelvtudományi Intézete, Bp. 5–17.

BALOGH LAJOS – KISS GÁBOR 2002. Analitikus elemzési lehetőségek A magyar nyelvjárások atlaszának számítógépre vitt anyagán. In: SZABÓ és mtsai. szerk. 2002: 41–54.

BENKŐ LORÁND 1967. A nyelvföldrajz történeti tanulságai. A Magyar Tudományos Akadémia Nyelv- és Irodalomtudományi Osztályának Közleményei 24: 29–48.

BENKŐ LORÁND 1989. A csángók eredete és települése a nyelvtudomány szemszögéből. Magyar Nyelv 271–87, 385–405.

BODÓ CSANÁD 2007. Követéses geolingvisztikai vizsgálat Moldvában. In: GUTTMANN–MOLNÁR szerk. 2007: 37–47.

BODÓ CSANÁD – VARGHA FRUZSINA SÁRA 2007. Jelenségtérképek A moldvai csángó nyelvjárás atlaszából. CD-ROM. Eötvös Loránd Tudományegyetem, Magyar Nyelvtörténeti, Szociolingvisztikai, Dialektológiai Tanszék, Bp.

CHAMBERS, JACK K. – TRUDGILL, PETER 2004.[2] Dialectology. Cambridge University Press, Cambridge.

GOOSKENS, CHARLOTTE – WILBERT HEERINGA 2004. Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. Language Variation and Change 16: 189–207.

GUTTMANN MIKLÓS – MOLNÁR ZOLTÁN szerk. 2007. V. Dialektológiai Szimpozion. BDTF Magyar Nyelvészeti Tanszéke, Szombathely.

HEGEDŰS ANDREA 2002. Nyelvjárási atlaszok egyesítési lehetőségei. In: SZABÓ és mtsai. szerk. 2002: 120–7.

HEGEDŰS ANDREA 2007. A kétféle *é* fonéma megléte és kiterjedtségének mértéke a Szilágyságban. In: GUTTMANN–MOLNÁR szerk. 2007: 120–5.

JUHÁSZ DEZSŐ 2004. Az északi csángók eredetéről „A romániai magyar nyelvjárások atlasza" tükrében. In: P. LAKATOS ILONA – T. KÁROLYI MARGIT szerk., Nyelvvesztés, nyelvjárásvesztés, nyelvcsere. Tinta Könyvkiadó, Bp. 164–74.

KIRÁLY LAJOS 2005. Somogy–zalai nyelvatlasz. A Magyar Nyelvtudományi Társaság Kiadványai 223. Bp.

KISS JENŐ 2002. Gondolatok a magyar dialektológiáról az új évezred küszöbén. In: SZABÓ és mtsai. szerk. 2002: 11–8.

KISS JENŐ szerk. 2004. Nyelv és nyelvhasználat a moldvai csángók körében. A Magyar Nyelvtudományi Társaság Kiadványai 222. Bp.

KRETZSCHMAR, WILLIAM 2001. Linguistic Databases of the American Linguistic Atlas Project (ALAP). http://www.ldc.upenn.edu/annotation/database/ papers/Kretzschmar/11.3.kretzschmar.pdf.

MÁRTON GYULA – HEGEDŰS ATTILA 2000. Szilágysági nyelvatlasz. Magyar Nyelvtudományi Társaság – Pázmány Péter Katolikus Egyetem, Bp.–Piliscsaba.

MCsNyA. = GÁLFFY MÓZES – MÁRTON GYULA – SZABÓ T. ATTILA szerk. 1991. A moldvai csángó nyelvjárás atlasza. I–II. kötet. A Magyar Nyelvtudományi Társaság Kiadványai 193. Bp.

MNyA. = DEME LÁSZLÓ – IMRE SAMU szerk. 1968–1977. A magyar nyelvjárások atlasza. I–VI. kötet. Akadémiai Kiadó, Bp.

CS. NAGY LAJOS 2004. Szóföldrajzi térképlapok a kolozsvári archívumban. Előadás a VII. Nemzetközi Magyar Nyelvtudományi Kongresszuson, Bp. Http://www.nytud.hu/NMNyK/eloadas/csnagy-ea.rtf.

NERBONNE, JOHN – WILBERT HEERINGA 2001. Computational Comparison and Classification of Dialects. Dialectologia et Geolinguistica 9: 69–83.

NERBONNE, JOHN – WILLIAM KRETZSCHMAR 2003. Introducing Computational Techniques in Dialectometry. In: NERBONNE–KRETZSCHMAR eds., Computational Methods in Dialectometry. Special issue of Computers and the Humanities, 37/3: 339–57.

PÉNTEK JÁNOS 2004. A moldvai magyar nyelv szótára – elvek és problémák. In: KISS szerk. 2004: 180–96.

PUTNAM, MICHAEL T. – D. CHRIS JOHNSON 2006. How far have we come and where do we still need to go? – The evolving dialect atlas. In: BEREND, NINA – ELISABETH KNIPF-KOMLÓSI szerk., Sprachinselwelten – The World of Language Islands. Peter Lang, Frankfurt am Main.

RMNyA. = MURÁDIN LÁSZLÓ gyűjt. – JUHÁSZ DESZŐ szerk. 1995–2004. A romániai magyar nyelvjárások atlasza I–IX. kötet. Magyar Nyelvtudományi Társaság, Bp.

SAJTOS JÓZSEF 2004. Sárvíz menti nyelvatlasz. A Magyar Nyelvtudományi Társaság Kiadványai 220. Bp.

SÁNDOR ANNA 2004. A Nyitra-vidéki magyar nyelvjárások atlasza. Kalligram, Pozsony.

SZABÓ GÉZA – MOLNÁR ZOLTÁN – GUTTMANN MIKLÓS szerk. 2002. IV. Dialektológiai Szimpozion. BDTF Magyar Nyelvészeti Tanszéke, Szombathely.

SZABÓ T. ATTILA 1959. A moldvai csángó nyelvjárás kutatása. Magyar Nyelvjárások 5: 3–38.

SzNySz. = GÁLFFY MÓZES – MÁRTON GYULA 1987. Székely nyelvföldrajzi szótár. Akadémiai Kiadó, Bp.

TÁNCZOS VILMOS 2004. A moldvai csángók nyelvészeti kutatása (1945–2004). In: KISS szerk. 2004: 208–85.

VÉKÁS DOMOKOS 1999. Informatikai lehetőségek a dialektológiában, különös tekintettel a fonetikai szempontokra. Kézirat. Http://www.btk.elte.hu/fonetika/p/s/fonszem.html.

VÉKÁS DOMOKOS 2000. Magánhangzó-rendszerek elemzése informatizált nyelvjárási korpuszon. In: GÓSY MÁRIA szerk., Beszédkutatás 2000. Beszéd és társadalom. MTA Nyelvtudományi Intézete, Bp. 75–86.

BODÓ CSANÁD – VARGHA FRUZSINA SÁRA