Vargha Fruzsina Sára – Vékás Domokos

Lokalizálható nyelvtörténeti adatok informatizálása és térképezése

A magyar nyelvjárási atlaszok és nyelvföldrajzi szótárak nélkülözhetetlen forrásai a magyar nyelvvel, kultúrával foglalkozó kutatóknak. Ezek az adattárak azonban eredeti formájukban nehezen kutathatók, hiszen több száz, esetleg több ezer nyomtatott térképlapból, illetve szócikkből állnak, így igen időigényes megtalálni, összegyűjteni és csoportosítani a vizsgálati szempontok szerint releváns adatokat. Több adattár egyidejű vizsgálata hagyományos módszerekkel különösen nagy kihívást jelent. A nyelvjárási adattárak elemzésének megkönnyítése, ezáltal a bennük rejlő forrásanyag valorizálása volt az elsődleges célja a Bihalbocs néven ismertté vált nyelvészeti technológiák fejlesztésének és a nyelvjárási adattárak informatizálásának (honlap: www.bihalbocs.hu).[1]

A nyelvjárási adattárak számítógépes feldolgozásához kifejlesztett nyelvészeti technológiák felhasználásával kezdtük el 2006-ban Szabó T. Attila Erdélyi Helynévtörténeti adattárának számítógépes feldolgozását. A nyomtatásban már megjelent adatok informatizálása a háromszéki történeti helyneveket tartalmazó kötettel[2] kezdődött. Az adatok denotátumfajták szerinti annotálásához, majd adatbázisba rendezéséhez, kereséséhez és térképezéséhez szükséges célprogramot fejlesztettünk ki (Olló néven) a dialektológiai megoldások adaptálásával és Hajdú Mihály módszertani észrevételeinek figyelembevételével.[3] A denotátumfajták szerinti annotálást kezdetben Bárth János végezte, majd a későbbiekben is ő irányította ezt a munkafázist.[4]

Történeti szövegekről lévén szó, a számítógépes feldolgozás során a legelső probléma, amit feltétlenül meg kellett oldanunk, a történeti grafémák (karakterek) kódolása volt. Ehhez a magyar egyezményes hangjelölési rendszer számítógépes alkalmazásában összegyűlt tapasztalatokból kiindulva egy analitikus kódrendszert dolgoztunk ki.[5] A mostani magyar ábécében meglévő ékezetes betűket megtartottuk, az összes többi, a történeti szövegekben előforduló ékezetes betűt mozaikszerűen építjük föl, külön az alapkarakterre illesztve a szükséges ékezet(ek)et (lásd az *1. ábrát*). Maga a rendszer a kötetek informatizálásának előrehaladtával folyamatosan bővül, minden kötetben akad egy-egy olyan graféma, amely korábban még nem fordult elő. Az adatok megfelelő kódolásához pedig minden egyes, az adattárban előforduló

Vargha Fruzsina Sára (1979) – tudományos munkatárs, PhD, ELTE, Budapest, fruzsa@gmail.com
Vékás Domokos (1962) – PhD, bihalbocs@gmail.com

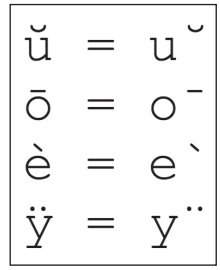
[1] Vékás Domokos: *Számítógépes dialektológia = V. Dialektológiai Szimpozion.* Szerk. Guttmann Miklós – Molnár Zoltán. Szombathely 2007. 289–293.

[2] *Szabó T. Attila erdélyi történeti helynévgyűjtése 2. Háromszék.* Szerk. Hajdú Mihály – Slíz Mariann. Budapest. 2001.

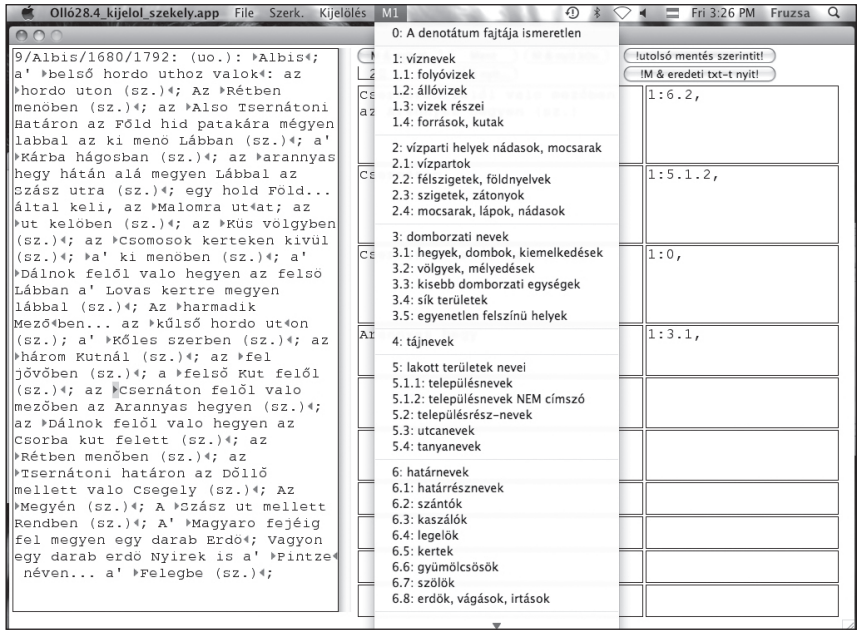
[3] Vargha Fruzsina Sára: *Nyelvjárási és helynévtörténeti anyagok számítógépes feldolgozása*. = *Kontextus – Filológia – Kultúra. II.* Szerk. František Alabán. Besztercebánya–Eger 2008. 77-84.

[4] Bárth M. János: *Háromszéki helynevek nyelvészeti elemzése informatikai módszerekkel*. Helynévtörténeti Tanulmányok II(2006). 207–217.

[5] A szövegek megjelenítéséhez szükséges betűkészlet és kódrendszer kialakításához Korompay Klárától kértünk és kaptunk segítséget, útmutatást, amelyet itt is szeretnénk megköszönni.



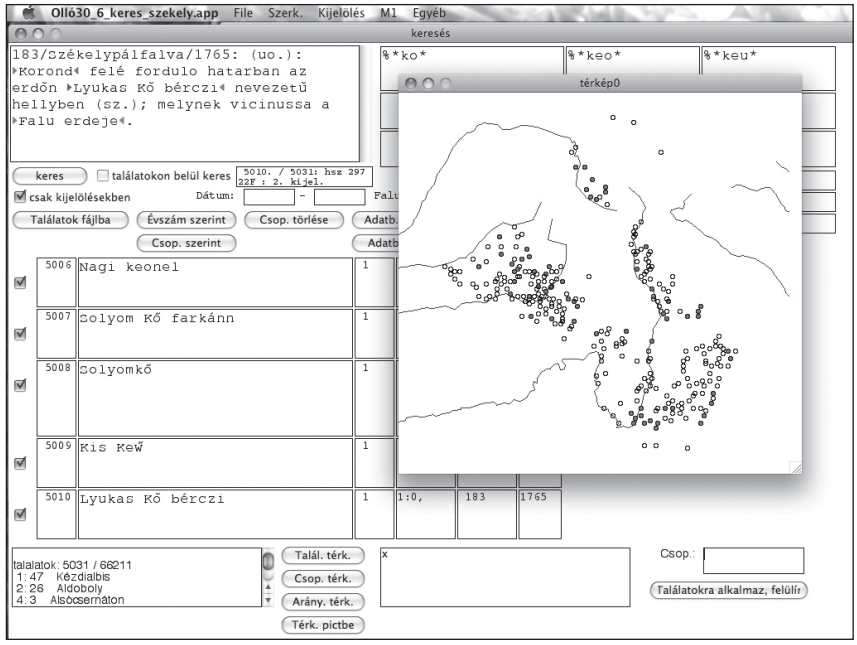
1. ábra. **A történeti karakterek analitikus kódolása**



2. ábra. **Helynevek kijelölése és denotátumfajták szerinti annotálása az Ollóban**

karaktert számon kell tartanunk (kódolnunk kell), hogy az informatizáláshoz szükséges konverzió után pontosan úgy láthassuk viszont, ahogyan Szabó T. Attila annak idején cédulára írta.

Eddig összesen nyolc kötet anyagát dolgoztuk föl a fent leírt kódolási rendszernek megfelelően, a nyomtatott változathoz készített elektronikus dokumentumokból kiindulva: 1. Alsófehér megye. Közzéteszi Hajdú Mihály és Janitsek Jenő. 2001. 204 lap. – 2. Háromszék. Közzéteszi Hajdú Mihály és Slíz Mariann. 2001. 207 lap. – 3. Szilágy megye. Közzéteszi Hajdú Mihály és Sebestyén Zsolt. 2002. 247 lap. – 4. Kisküküllő és Nagyküküllő megye. Közzéteszi Hajdú Mihály és Sebestyén Zsolt. 2003. 272 lap. – 5. Torda-Aranyos megye. Közzéteszi Hajdú Mihály, Buboly Magdolna és Sebestyén Zsolt. 2004. 252 lap. – 6. Udvarhelyszék. Közzéteszi Hajdú Mihály és Bárth M. János. 2005. 254 lap. – 7. Maros-Torda



3. ábra. **Keresés az Ollóban a *kő* különböző írásmódú változataira**

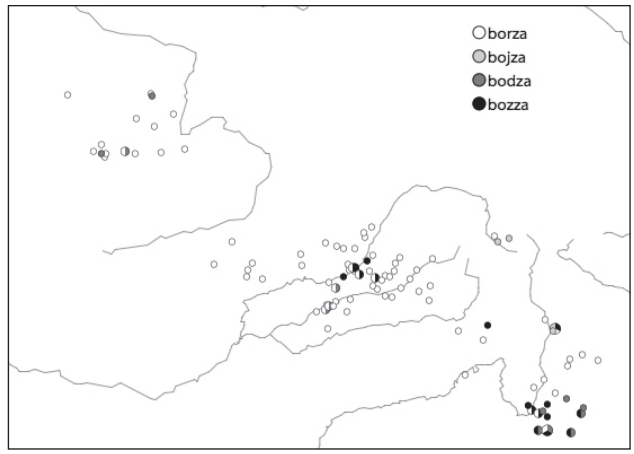
megye. Közzéteszi Hajdú Mihály és Sófalvi Krisztina. 2005. 812 lap; két részkötetben: A. 1–413, B. 415–812. – 8. Csík-, Gyergyó- és Kászonszék. Közzéteszi Hajdú Mihály, Makay Emese és Slíz Mariann. 2006. 153 lap.

Az informatizált változat alapegysége az eredeti, Szabó T. Attila által lejegyzett céduláknak felel meg: egy adott településről, egy forrásból, egy adott időpontból származó adatok összességét tekintjük feldolgozási alapegységnek. Az így kialakított rendszer tehát pontosan leképezi a forrásdokumentumét (akárcsak a nyelvjárási adattárak feldolgozása esetében), ugyanakkor megőrzi a kapcsolatot a nyomtatott változattal is az oldalszámok adatokhoz kapcsolásával.

A névtani szempontú feldolgozás erre a célra kialakított speciális környezetben történik (*2. ábra*), ahol a névtesteket, helynevet tartalmazó körülírásokat manuálisan, az egérrel, illetve speciális billentyű- és menüparancsok segítségével lehet a szövegben kijelölni, további részekre bontani, az egyes kijelöléseket, kijelölésrészleteket előre kialakított tipológiák szerint minősíteni.[6]

A feldolgozás során kijelölt és minősített helynevek, helyneveket tartalmazó körülírások adatbázisba rendezésük után az Olló keresőfelületén különféle szempontok szerint lekérdezhetők, csoportosíthatók, térképezhetők. Az adatokban való keresést a *3. ábra* szemlélteti. Az ábrán látható példán a *kő* szó előfordulásait kérdeztük le az elsőként informatizált székelyföldi adatokban. Az Olló program az adatok és a keresett szó írásmódjának egyszerűsítésével segíthet minket abban, hogy írásmódtól függetlenül megtaláljunk egy szót, szókapcsolatot. Így a *kő* szó valamennyi előforduló változatát megkapjuk, ha rákeresünk a *ko, keo* és *keu* betűkapcsolatokra. (A keresett kifejezések

[6] Bárth M. János: *Székelyföldi történeti helynevek nyelvi elemzése*. Doktori értekezés. ELTE BTK. Bp. 2010. 26–40.

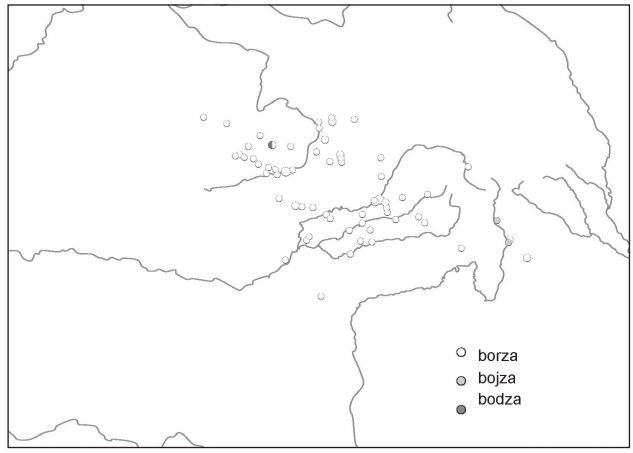


***4. ábra.* A *bodza* alakváltozatainak területi megoszlása a helynévtörténeti adatokban**

***5. ábra.* A *bodza* legkorábban előforduló (17. századi) változatai a helynévtörténeti adatokban és térbeli elhelyezkedésük**

megadására szolgáló mezők jobbra fönt, a találatok a kép bal oldalán, egymás alatt láthatók.) A keresés eredményeképpen 5031 találatunk lesz. Az adatok csoportosítása révén lehetőségünk van azoknak az adatoknak a kiválogatására, amelyek valóban a *kő* szót tartalmazzák. (Hiszen a keresésnek megfelelő betűkapcsolatok más szavakban is előfordulhatnak, pl. *Kőles szer, Lökös, Keozepseo hatar.*) A kiválogatott adatok térbeli elhelyezkedését azonnal térképre is vetíthetjük, és lehetőségünk van arra is, hogy az adatokat tetszőleges szempontok szerint tovább csoportosítsuk.

A *4. ábra* a *bodza* különböző alakváltozatainak előfordulását mutatja a helynévtörténeti adatokban. A településeket jelző karikák mérete az egy-egy településről származó adatok mennyiségével arányosan változik. A legáltalánosabban elterjedt változat a *borza,* de



***6. ábra.* A *bodza* alakváltozatai az Erdélyi magyar szótörténeti tár adataiban**

jellemző területi kötöttséggel megjelenik a *bozza,* a *bodza* és a *bojza* alak is. Mivel ismerjük az adatok keletkezésének idejét, rendezhetjük az adatainkat évszám szerint, és így térképezhetjük a legkorábbi, 16. századi adatokat (*5. ábra*). Igen kevés adatunk van ebből a korai időszakból, de a kirajzolódó térkép alapján mégis érdekes felfedezést tehetünk. A legkorábbi történeti helynévadatokban Borzaszeg neve kizárólag *Bozzaszeg* változatban fordul elő a településen és környékén. Egészen az 1700-as évekig a *bozza* a jellemző változat a környék helyneveiben, de később teljesen eltűnik, és az 1700-as évek végén már csak a Székelyföldön és Háromszéktől délre találunk a *borzá*tól eltérő alakváltozatokat.[7]

Az *Erdélyi magyar szótörténeti tár*[8] adatai többnyire épp úgy lokalizálhatók a forrás alapján, akár a történeti helynévadatok, nem lehetetlen vállalkozás tehát, hogy a szótár adatait térben elhelyezve történeti nyelvjárási térképeket hozzunk létre. Míg azonban az informatizált, adatbázisba rendezett helynévanyag kereshető és automatikusan térképezhető, addig a Tár egyelőre szerkesztett könyvként áll rendelkezésünkre. A közelmúltban elkészült PDF-változat némely tekintetben megkönnyíti a címszók böngészését, noha nyilvánvalóan nem nyújt olyan sokrétű lehetőségeket a kutató számára, mint az EHA 2006 óta nyolckötetnyire duzzadt adatbázisa.

A *6. ábrán* a *bodza* különböző alakváltozatait térképeztük a Tár adatai alapján. A történeti helynevekhez képest azonban ezen a térképen szinte egyeduralkodó a *borza* változat, ezen kívül csak két *bojza* és egy *bodza* adatot találunk, mindhármat a szócikk végén, a helyneveknél.

A *bodza* térképezése a két adattár alapján egyrészt szemlélteti a lokalizálható történeti adatok ilyetén feldolgozásában rejlő kutatói lehetőségeket, másrészt rámutat arra, hogy a két adattár szervesen összefügg, a bennük rejlő adatok a Tár megfelelő feldolgozása és a két adatbázis integrálása révén jól kiegészíthetnék egymást.

[7] A bodza alakváltozatainak területi megoszlásáról és annak tanulságairól lásd bővebben Vargha Fruzsina Sára: *A dialektometria alkalmazása és történeti helynevek nyelvföldrajzi vizsgálata a Székelyföldön*. Helynévtörténeti Tanulmányok V(2010). 223–233.

[8] *Erdélyi magyar szótörténeti tár.* I–XII. Szerk. Szabó T. Attila és munkatársai. Buk.–Bp.–Kvár 1979–2009.

Digitization of Transylvanian Historical Place-Names

Keywords: Transylvanian historical place-names, Attila Szabó T., Transylvanian Hungarian Historical Thesaurus

In 2010 the last volume of the Transylvanian Historical Place-Names series collected by Attila Szabó T. was published. Nevertheless, in its original, paper-based form the corpus of approximately 600 000 place-names is difficult to study as it is not searchable. Based on the linguistic technologies used in previous projects aiming the digitization of Hungarian dialect data, a method and a related software tool have been developed for the digitization of the historical place-names. In the database the place-names are classified according to a predefined list of denotation types. A special encoding system provides the possibility to make searchable the historical data while maintaining its original form (special characters and diacritics). As every item belongs to a location, they can be easily represented on maps. Localized data coming from an other but highly related corpus, the Transylvanian Hungarian Historical Thesaurus (Erdélyi magyar szótörténeti tár) can also be mapped. After the digitization of the Thesaurus, the two collections shall ideally complete each other as presented here with the mapping of the different forms of the word *bodza* (‘elder’).