MAGYAR NYELVJÁRÁSOK 53 (2015): 47–61.
A DEBRECENI EGYETEM MAGYAR NYELVTUDOMÁNYI TANSZÉKÉNEK LEKTORÁLT FOLYÓIRATA

Névföldrajzi térképlapok klaszteranalízise Zala megye helynevei alapján*

DITRÓI ESZTER

1. Bevezetés

A helynevek rendszerszerűségét és mintákhoz való igazodásukat a névkutatók jó ideje általánosan vallják. Ez a gondolat nyújt kiindulási alapot ahhoz a felvetéshez is, hogy egy-egy táj névadási mintáinak feltérképezése és azok összevetése által kirajzolódhatnak egységes névadási sajátságokkal rendelkező névföldrajzi tájak, azaz meghatározhatók egyes ún. névjárási területek (erre vonatkozóan lásd pl. DITRÓI 2010, 2011, 2012). Minthogy a névrendszerek területi különbségeinek a megragadására megítélésem szerint a statisztikai összevető módszerek nyújtják — mivel objektív, számszerűsíthető eredményekkel dolgoznak — a legbiztosabb alapot, vizsgálataimat egy mátrix alapú metódusra, valamint a kialakított mátrix dimenziócsökkentő eljárására (klaszteranalízis) alapozom, s ezáltal igyekszem egy megyényi térségre vonatkozóan, konkrétan Zala megye területén a névföldrajzi régiókat feltérképezni.

2. Szakirodalmi és módszertani áttekintés

A címben megjelölt témával összefüggésben a tudománytörténeti előzményeket és a módszertani alapvetést elsősorban két tudományterület, a névföldrajz és a statisztika kapcsolatára építve érdemes részletezően bemutatni.

Az a felfogás, miszerint bizonyos tájak helynévkincse rendszert alkot, azaz a helynévrendszer tanulmányozásakor valójában egyes vidékek névrendszeréről beszélhetünk, jó ideje egyöntetű álláspontnak tekinthető a névtani szakirodalomban (lásd például HOFFMANN 1993: 29, 64, 82–83, 85, 87, HAJDÚ 1991, 1999, TÓTH 2001: 222, VARGHA 2010, BÁRTH 2010). E témakör matematikai statisztikai alapú megközelítése azonban meglehetősen újkeletű, ezért érdemes részletesebben is ismertetni e módszertani eljárás lényegét. Ezt úgy igyekszem az alábbiakban megtenni, hogy konkrétan a helynévrendszerekre vonatkoztatom minden egyes jellemzőjét.

* A publikáció az MTA–DE Magyar Nyelv- és Névtörténeti Kutatócsoport programja keretében készült.

A helynévrendszerek statisztikai megközelítéseinek alapja maga a meghatározott szempontrendszer szerint kiválasztott névanyag, amelynek névrendszertani elemzését a HOFFMANN ISTVÁN által kidolgozott helynévelemzési modell (1993) szerint végeztem el. Az ilyen jellegű analízisnél fontos, hogy a névanyag vizsgálatát településenként külön-külön valósítsuk meg — annak az általános feltevésnek megfelelően, hogy egy-egy (főleg kisebb) település helynevei bizonyos tekintetben rendszert alkotnak —, ez biztosítja ugyanis a névrendszerek összevethetőségét. Az elemzés során kapott helynévszerkezeti típusokat ezt követően gyakorisági sorokba rendezzük, melyek kialakítása helynévszerkezeti kategóriák szerint történik.[1] Ha pedig ezeket az adatokat térképre vetítjük, jól kirajzolódik, hogy az egyes települések helynévanyaga mely más településekével mutat hasonlóságot, s melyekétől különbözik jelentősebb mértékben. Azt is hangsúlyoznunk kell ugyanakkor, hogy a települések hasonlósági mintái nem feltétlenül mutatnak egységes képet. Ennek a problémának a megoldására alkalmazzák a statisztikában a klaszteranalízist, amelynek révén a mátrix alapján kapott eredmények egyetlen térképlapon ábrázolhatók, ezáltal pedig könnyebben megragadhatók a különböző névföldrajzi területek.

A klaszteranalízis gyökerei igen korai időkre nyúlnak vissza. Minthogy alapvetően egy csoportosító eljárásról van szó, csírái már az ókorban is jelen voltak. A klaszteranalízis elterjedésében a 18. században a svéd LINNÉ által megalkotott állat- és növényrendszertan jelentette a mérföldkövet. Az ilyesfajta csoportosító módszereket eleinte csak a biológia területén alkalmazták, és csupán később történtek kísérletek arra, hogy más területekre is alkalmazzák az eljárást (vö. JARDINE–SIBSON 1971). A nyelvészetben is használták például a különböző nyelvek rokonsági fokának a meghatározására (vö. ehhez ROBINS 1999: 181–206). A módszer idővel nagyfokú változáson ment keresztül köszönhetően a matematikai statisztika új eredményeinek és a számítógépes technológiának.

A klaszteranalízist a szakirodalomban általában kétféle megközelítésben találjuk meg: egyes tanulmányok a módszer matematikai leírását, a különféle klaszterező eljárások és távolságszámítási módok részletes leírását tűzik ki célul. A módszert azonban alkalmazói oldalról is bemutathatjuk: e megközelítések általában problémafelvető esettanulmányok formájában jelentkeznek. Az első típusba tartozó, a matematikai alapokat ismertető művek zömmel az 1970-es évektől jelentek meg, közöttük említhető meg ANDERBERG (1973) vagy EVERITT (1974) monográfiája, illetve a magyar szakirodalomban például PÁRNICZKY GÁBOR publikált összefoglaló munkát a klaszteranalízisről (1976). Az alkalmazói oldalt

[1] A gyakorisági sorok alapján aztán létrehozunk egy mátrixot. Több lehetőség is rendelkezésünkre áll, én magam a BRAY–CURTIS-féle metódust (1957) alkalmaztam jelen munkámban, mely 0–1-ig tartó értékben jeleníti meg két település hasonlóságát. Ezekről a módszertani lépésekről, illetve a mátrix kialakításáról részletesebben lásd DITRÓI 2015.

szem előtt tartó munkákat a természettudományoktól a társadalomtudományokig igen sokféle tudományterület metodológiai eljárásai között ott találjuk. A nyelvészeten belül leginkább a dialektológia mutatott ezidáig nagyobb fokú érdeklődést az új típusú klaszterezési eljárások iránt (vö. ehhez pl. VARGHA–BODÓ–VÉKÁS 2012, VARGHA–VÉKÁS 2012).

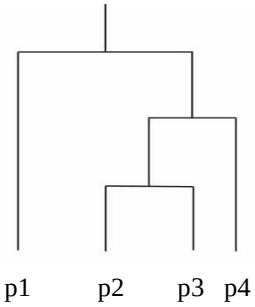
Klaszteren, azaz csoporton elemek együttesét értjük, amelyeket egy jól definiált szempont szerint hasonlónak tekintünk. A klaszteranalízis tehát egy típusalkotó eljárás, amellyel az elemeket, egyedeket homogén csoportokba soroljuk (BOLLA–KRÁMLI 2012: 313). A klaszteranalízisnek két általános eljárását különíthetjük el: beszélhetünk hierarchikus és nem hierarchikus elemzési módszerekről (vö. ehhez i. m. 313–316). A hierarchikus klaszterezésnél nem kell előre ismernünk a létrehozandó klaszterek számát, míg a nem hierarchikusnál már a kiinduláskor meg kell adnunk a lehetséges csoportokat. Minthogy a névrendszerekre a hierarchikus eljárásokat alkalmazom, ezekről részletesebben is célszerű szólnunk.

A hierarchikus eljárásoknak két altípusát említhetjük meg: az egyiket agglomeratív, a másikat divizív eljárásnak nevezzük. Az agglomeratív eljárásban a kiinduláskor több klaszterünk van, majd lépésről lépésre csökkentjük a klaszterek számát egészen addig, amíg az utolsó lépésnél már csak egyetlen klaszterünk marad. Az eljárás abban a tekintetben is hierarchikus, hogy egyszerre mindig csak két klaszter egyesíthető, és azok aztán az elemzés végéig együtt értelmezendők. A helynévrendszerek kapcsán, tehát a névkutatásban ez úgy értelmezhető, hogy vesszük a különböző települések névrendszereit, s azokat egyre nagyobb csoportba soroljuk a hasonlóságuk alapján: mindig a kisebb csoport felől haladunk tehát a nagyobb felé. Ez a módszer empirikus-induktív sajátságokkal jellemezhető. A divizív eljárás ennek a fordítottját jelenti: itt a kiinduláskor csupán egyetlen klaszterünk van, majd lépésenként különválasztjuk őket: azaz az összes, általunk elemzett település névrendszerét egy csoportnak tekintjük, s ezen belül alakítunk ki különböző alcsoportokat. Minthogy a legtöbb klaszterezési eljárás agglomeratív módszereket alkalmaz, én magam is ezt használom fel a névrendszerek elemzésekor.

A hierarchikus eljárások az adatelemeket úgynevezett fagráfba, dendogramba rendezik. Ez az eljárás nem teljesen új a nyelvészetben sem: az egyes nyelvcsaládokhoz tartozó nyelvek rokonsági fokának érzékeltetésére használt családfamodellek például egy ehhez nagyban hasonló módszert követnek. A klaszteranalízis során létrehozott dendogramok azonban abban térnek el ettől az összehasonlító nyelvészetben már meghonosodott módszertől, hogy a fák ágai itt rögzítettek, el nem mozdíthatók, s a csoportok nem cserélhetők fel (lásd az 1. ábrát).

Az 1. ábrán látható dendogram a következőket mutatja: adott 4 település helynévrendszere (p1, p2, p3, p4), melyeket a mátrixból kapott hasonlósági adataik

segítségével csoportokba kívánunk sorolni. Látható, hogy a p2 és p3 jelzésű települések névrendszere nagyobb hasonlóságot mutat, ezért azokat egy csoportba sorolta a dendogram. Tovább haladva a p4-es jelzetű településről is megállapítható, hogy távolabbi kapcsolattal ugyan, de a p2–p3 csoporttal egy típusba tartozik. A p1 pedig a többi névrendszerhez képest különálló jegyeket mutat. Mindez tehát azt jelenti, hogy egy közös pontból kiindulva két nagyobb csoportot határolhatunk körül, majd ahogy szigorítjuk a hasonlósági fokot, további három alcsoportot tudunk elkülöníteni. Ennek a módszernek több település névrendszerére való kiterjesztésével véleményem szerint jól elkülöníthetők az egymáshoz hasonló helynévrendszertani jegyeket felmutató települések, s ezáltal pedig maguk az egyes névföldrajzi területek.

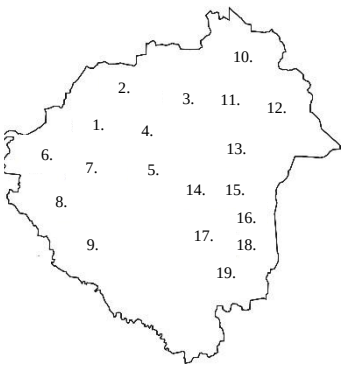


1. ábra: Dendogram

3. A kutatási terv

Kutatási tervem alapvetően négy pilléren nyugszik: a mintavételen, az ún. gyakorisági sorok kialakításán, a mátrix, valamint a klaszteranalízis során kapott eredmények értelmezésén. A mintavétel során a Magyar nyelvjárások atlasza kutatópontjaihoz igazodtam, így a vizsgálathoz kiválasztott Zala megyében 19 kutatópont mikrotoponimáit elemeztem. Ezeket a 2. ábra szemlélteti.

1. Csöde
2. Zalaháshágy
3. Ságod
4. Babosdöbréte
5. Zalatárnok
6. Szentgyörgyvölgy
7. Kerkabarabás
8. Bödeháza
9. Kerkateskánd



10. Pakod
11. Padár
12. Zalacsány
13. Szentpéterúr
14. Hahót
15. Egeraracsa
16. Balatonmagyaród
17. Gelsesziget
18. Galambok
19. Pat

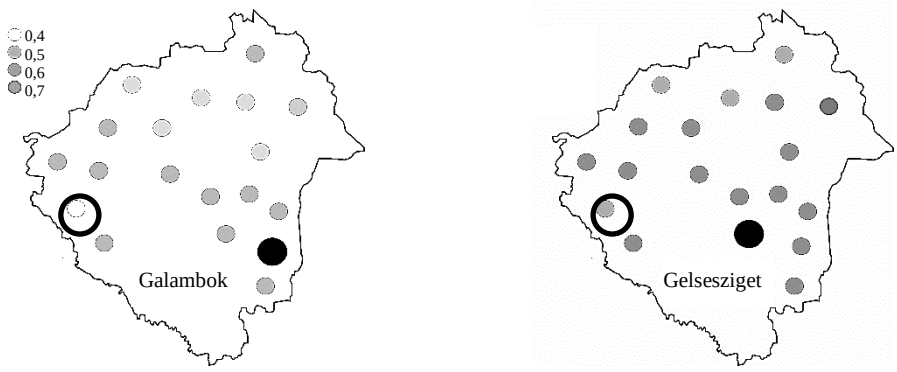
2. ábra: Zala megyei kutatópontok

Az MNyA. kutatópontjaihoz való alkalmazkodást az indokolja, hogy ezáltal statisztikailag is összevethető térképeket készíthetünk a dialektometria által létrehozott nyelvföldrajzi térképekkel. Ennek a döntésnek ugyanakkor ezzel az előnnyel szemben a hátránya az, hogy a mintánk nem tudja a teljes Zala megye helynévrendszereinek jellegzetességeit bemutatni.

A gyakorisági sorok felállítását a HOFFMANN ISTVÁN-féle helynévelemzési modell (1993) kategóriarendszeréhez igazodva végeztem el, a mátrix kialakítása pedig a BRAY–CURTIS-féle metódus (1957) szerint történt. A klaszteranalízis során a kialakított többdimenziós mátrixunk eredményeit osztályokba soroljuk, így kialakítva egy dendogramot (fagráfot), amelynek a segítségével megállapíthatjuk, mely névadási területek mutatnak szorosabb, s melyek lazább kapcsolatot. A mostani tanulmányomban a klaszteranalízis különböző módszertani eljárásait mutatom be: a teljes lánc módszert, a csoportátlag módszert, a súlypont vagy centroid eljárást, valamint a variancia módszert (Ward metódust), s ezek felhasználási lehetőségeit vázolom fel a névrendszerek területi különbségeinek feltérképezésében. Az analízis végén aztán reményeim szerint az is kirajzolódik, hogy a lehetséges módszerek közül melyik lehet az itt vizsgált problémakör feltárásában a leginkább célravezető.

4. Eredmények

4.1. A jelenkori és régebbi névrendszerek területi differenciáltságára már az eddigiekben is történt jó néhány utalás a névtani szakirodalomban (a teljesség igénye nélkül lásd pl. KÁLMÁN 1967, TÓTH 1998: 121–134, 2002: 127–138, BÁBA 2013: 53, DITRÓI 2011, 2012, 2013, 2015), s ezek mindegyike arra mutatott rá, hogy ezek az eltérések a névadás egyes jellegzetességeinek a feltárása révén írhatók le. A Zala megye fent említett településeinek helynévanyagán végzett analízis ezt a megállapítást szintén megerősíti: a gyakorisági sorok alapján készített mátrix eredményeinek térképre vetítése során a megye északi és déli területeinek különbsége körvonalazódik.



3. ábra: Galambok és Gelsesziget hasonlósági térképlapja

Galambok névanyaga a megye névanyagával átlagosan 0,6-os egyezést mutat, azonban a terület északi részén, a Zala folyó egy szakaszán eltérő névadási mintákat tapasztalunk. Gelsesziget térképlapján hasonlóképpen regisztrálható ez a fajta különbség, azonban a névadási minták eltérő viselkedése e település esetében kissé északabbra regisztrálható. Ezenkívül a Kerka mellett elhelyezkedő Bödeháza (bekarikázva) névmintája mindkét bázistelepüléstől eltérést mutat. Az egyes települések mátrixa (a 3. ábrán bemutatottakhoz hasonlóan) tehát — noha bizonyos fokú hasonlóságot megfigyelhetünk — nem ad kiemelkedő egyezést. Ahhoz, hogy pontos képet tudjunk alkotni arról, hogy a kapott többdimenziós mátrix milyen valós névföldrajzi területeket jelöl ki, a klaszteranalízishez érdemes folyamodnunk.

4.2. A klaszteranalízis — ahogy a fentiekben már ismertettem — egy olyan dimenziócsökkentő eljárás, amelynek során a mátrix összes térképi eredményét egyetlen térképlapon tudjuk ábrázolni. Ez az eljárás kifejezetten akkor előnyös, ha a mátrix településenkénti hasonlósági eredményei — mint itt is — a vártnál gyengébb egyezést adnak. Zala megye 19 településének a vizsgálatakor ez azt jelenti, hogy noha körvonalazódni látszik egy északi, a Zala folyó mentén regisztrálható névadási mintákat érintő gát, a településenkénti eredmények ingadozása miatt mégsem tudjuk ezt teljes pontossággal meghatározni.

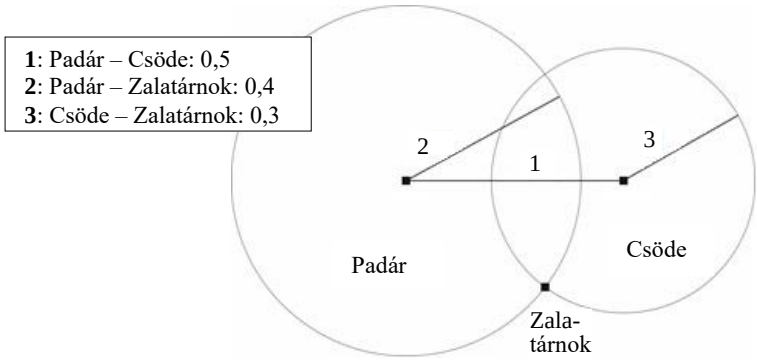
A klaszterezés egyik lényeges lépése az, hogy a mátrixból kapott adatokat térben ábrázolja. Ez egy kétdimenziós leképezés, amit a statisztikai programok önmaguk végeznek el, s amelynek szemléltetésére Padár, Csöde és Zalatárnok települések példáját használom fel. Az alábbi táblázatban e három település különbségmátrixát[2] láthatjuk.

	Padár	Csöde	Zalatárnok
Padár	0	0,5	0,4
Csöde	0,5	0	0,3
Zalatárnok	0,4	0,3	0

4. ábra: Padár, Csöde és Zalatárnok különbségmátrixa

Padár és Csöde névrendszereinek összevetése azt mutatja, hogy ezek 0,5-ös távolságra vannak egymástól, amit ha egy egyenesen veszünk fel, máris megkapjuk Padár és Csöde relatív helyzetét a kétdimenziós térben. Ezt követően Padár kiindulópontból felvesszük a Padár–Zalatárnok 0,4-es távolságot, s szerkesztünk egy 0,4 egységnyi sugarú kört. Ugyanezt tesszük Csöde esetében is: Csöde kiindulópontból felvesszük a Csöde–Zalatárnok 0,3-as távolságot, s szerkesztünk egy 0,3-as sugarú kört. Döntenünk kell arról, hogy felfelé vagy lefelé építjük a pontjainkat a térben. Ha lefelé, akkor a két kör alsó metszéspontjában helyezkedik el Zalatárnok. Ezzel el is készítettük a fenti három település kétdimenziós leképezését (lásd 4. ábra).

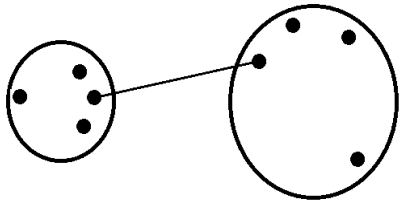
[2] A különbségmátrix a hasonlósági mátrix inverze. Tulajdonképpen a hasonlósági mátrix számadatait kivonjuk 1-ből, és máris megkapjuk a különbségmátrixot. Erre azért van szükség, mert a klaszteranalízis során a statisztikai programok döntő hányada különbségmátrixszal dolgozik.



5. ábra: Padár, Csöde és Zalatárnok kétdimenziós leképezése

A klaszterező eljárás során a program ezeket a pontokat klaszterekbe rendezi, s ezek pontjait összevetve megrajzolja a dendogramot. Amint azt a fentiekben már jeleztem, a klaszteregyesítő eljárások közül négyet, a teljes lánc módszert, a csoportátlag módszert, a centroid vagy súlypont eljárást és a variancia módszert (Ward módszert) használtam a névrendszerekre, de ezeken kívül még létezik egyszerű lánc eljárás is. Ezek jellemzőit az alábbiakban mutatom be.

4.2.1. Az egyszerű lánc módszer során a klaszterek hasonlósága a két különböző klaszter legközelebbi, azaz legjobban hasonlító elemén alapul. Ezt az eljárást úgy alakították ki, hogy lazán összetartozó, nagy csoportokat eredményezzen, emellett ez a módszer rendkívül érzékeny a kiugró értékekre és az esetleges hibás adatokra. Így akár bizonyos jelenségek vizsgálatára a névtani kutatásokban is alkalmazható, noha ennél jóval érzékenyebb összevető módszerek is léteznek. Az egyszerű lánc módszer sematikus ábrája látható alább: az ábrán szereplő pontok megrajzolása az 5. ábrán látható eljárással történik.



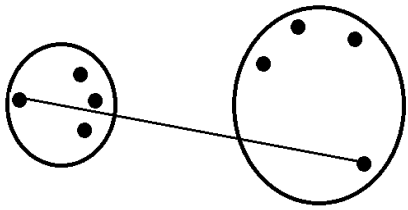
6. ábra: Az egyszerű lánc módszer sematikus ábrája

Kiugró értékek egy névrendszertani elemzés során többször is előfordulhatnak nyelvi és nyelven kívüli okok miatt egyaránt. Jól példázhatják ezt azok a névrendszerek, amelyek különböző tényezők folytán egyfajta névföldrajzi szigetként viselkednek.[3] Az egyszerű lánc módszer ezeknek a felderítésére alkalmas. Minthogy azonban nem minden adatot vet össze, csupán a legközelebbi pontokat veszi

[3] A névrendszerek területi differenciáltságára irányuló kutatások egyik izgalmas területe éppen az ilyen szigetek feltárása lehet.

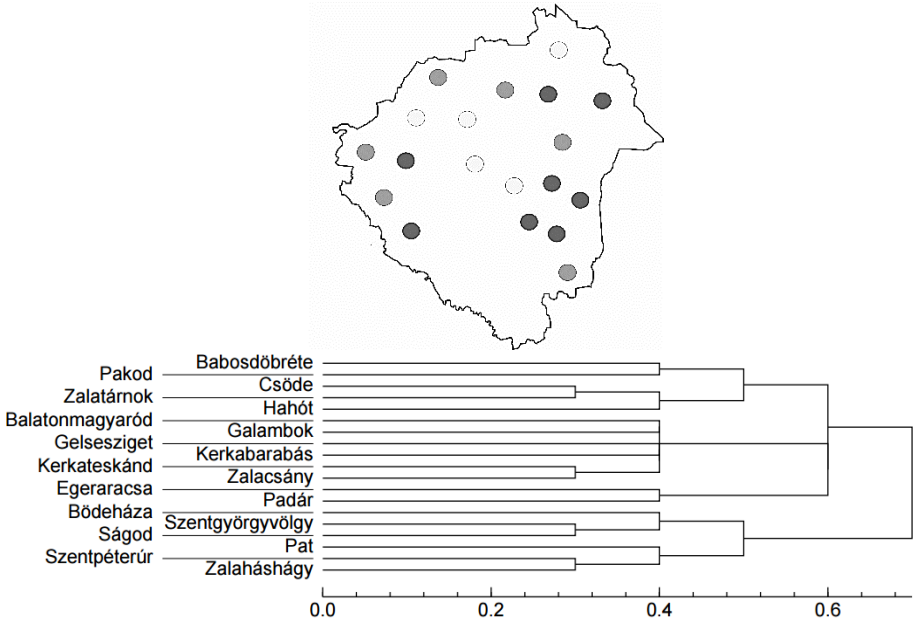
számításba, ez az eljárás kevésbé érzékeny volta miatt általában véve aligha lehet a névföldrajzi vizsgálatokhoz a legmegfelelőbb. (Az egyszerű lánc módszer további tulajdonságaihoz lásd TAN–STEINBACH–KUMAR 2006: 517.)

4.2.2. A teljes lánc módszer esetén a klaszterek hasonlósága a két klaszter legtávolabbi, azaz legkevésbé hasonlító elemén alapul. Erőssége ennek a módszernek az, hogy a kiugró és esetlegesen hibás adatokra kevésbé érzékeny, ugyanakkor nem képes kezelni az egyenlőtlen nagyságú klasztereket (lásd TAN–STEINBACH–KUMAR 2006: 517).



7. ábra: A teljes lánc módszer sematikus ábrája

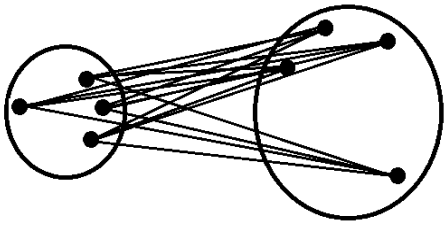
A módszer tehát eredendően nagy méretű és egyforma klaszterek kialakítására törekszik, ezáltal pedig kevésbé alkalmas a névrendszerekben körvonalazódó csoportok megrajzolására, hiszen az egyes névföldrajzi területek természetesen különböző kiterjedtségűek lehetnek. Zala megye 19 településén a teljes lánc módszerrel végzett analízis a következő eredményeket hozta.



8. ábra: A teljes lánc módszer térképe és dendogramja Zala megye 19 települése alapján

A fagráfon nem fedezhetünk fel egybefüggő névföldrajzi területeket. Ez nem is különösebben meglepő, s mivel ma már ismernek pontosabb klaszterezési eljárásokat is, ezt a módszert igen ritkán alkalmazzák a kutatásokban.

4.2.3. A csoportátlag módszer esetében két klaszter távolsága a klaszterekben előforduló elemek páronkénti távolságának az átlagán alapszik: azaz ebben az eljárásban minden elemet összevetünk minden elemmel.



9. ábra: A csoportátlag módszer sematikus ábrája

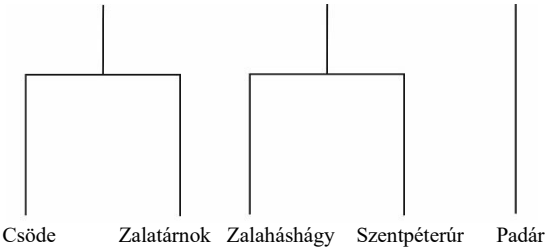
Előnye, hogy ez a módszer érzékeny a kiugró értékekre, s ezeket külön klaszterbe sorolja. Véleményem szerint ez a névrendszerek elemzésére vonatkozóan pozitív tulajdonság, hiszen általa még az esetleges szigetek, különleges névrendszertani területek is jól körvonalazhatók.

További előnyként emelhetjük ki még azt is, hogy a módszer algoritmusa is kifejezetten egyszerűnek mondható. Ennek szemléltetésére Padár, Csöde, Zalatárnok, Zalahásháhgy, valamint Szentpéterúr települések névrendszereit választottam ki. A gyakorisági soraik mátrix alapú értelmezése révén rájuk vonatkozóan a következő hasonlósági adatokat kaptuk.

	Padár	Csöde	Zalatárnok	Zalahásháhgy	Szentpéterúr
Padár	1	NA	NA	NA	NA
Csöde	0,5	1	NA	NA	NA
Zalatárnok	0,6	**0,7**	1	NA	NA
Zalahásháhgy	0,4	0,6	0,6	1	NA
Szentpéterúr	0,5	0,6	0,6	**0,7**	1

10. ábra: A gyakorisági sorok alapján létrehozott hasonlósági mátrix

A csoportátlag módszer azon alapul, hogy meg kell találnunk a legnagyobb hasonlóságot mutató adatokat. A mátrix alapján ez Zalatárnok–Csöde, valamint Szentpéterúr–Zalahásháhgy viszonylatában mutatkozik meg, mindkettő esetében 0,7-es hasonlósági fokot jelző adatot látunk (lásd a 11. ábrán). Így tehát az öt település névrendszerét három nagyobb csoportra bonthatjuk, s az eddigi eredmények tükrében az első dendogramjuk a következőképpen rajzolható meg (11. ábra).



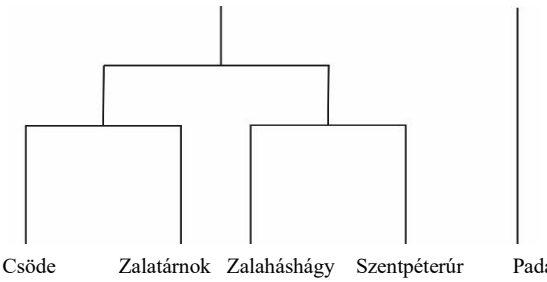
11. ábra: Az első dendogram kialakítása a mátrix alapján

Minthogy Csöde és Zalatárnok, illetve Zalahásháqy és Szentpéterúr már közös csoportot alkot, a mátrixban egyetlen közös cellában szerepeltethetjük őket, mégpedig a számadataik átlagát véve figyelembe. Így pedig a mátrix a 11. ábrán látottak szerint alakul.

	Padár	Csöde–Zalatárnok	Zalahás-hágy–Szent-péterúr
Padár	1	NA	NA
Csöde–Zalatárnok	(0,5+0,6)/2=0,55	1	NA
Zalahásháqy–Szentpéterúr	(0,4+0,5)/2=0,45	(0,6+0,6+0,6+0,6)/4=**0,6**	1

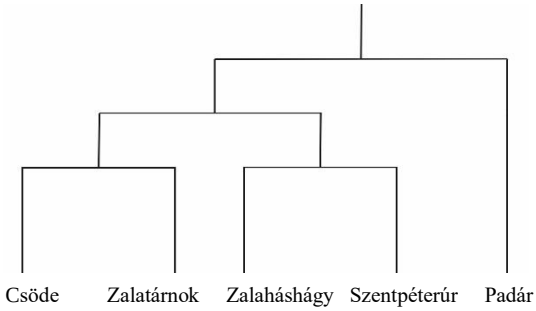
12. ábra: A csoportátlag módszeren alapuló klaszteranalízis első lépéseként létrehozott mátrix

Ezen a ponton ismét kiválasztjuk a legnagyobb értéket, ami a Csöde–Zalatárnok és a Zalahásháqy–Szentpéterúr közötti 0,6-os érték. Mindez pedig azt jelenti, hogy a következő nagyobb csoportot a második dendogramon Csöde–Zalatárnok, valamit Zalahásháqy–Szentpéterúr adja (lásd a 13. ábrán).



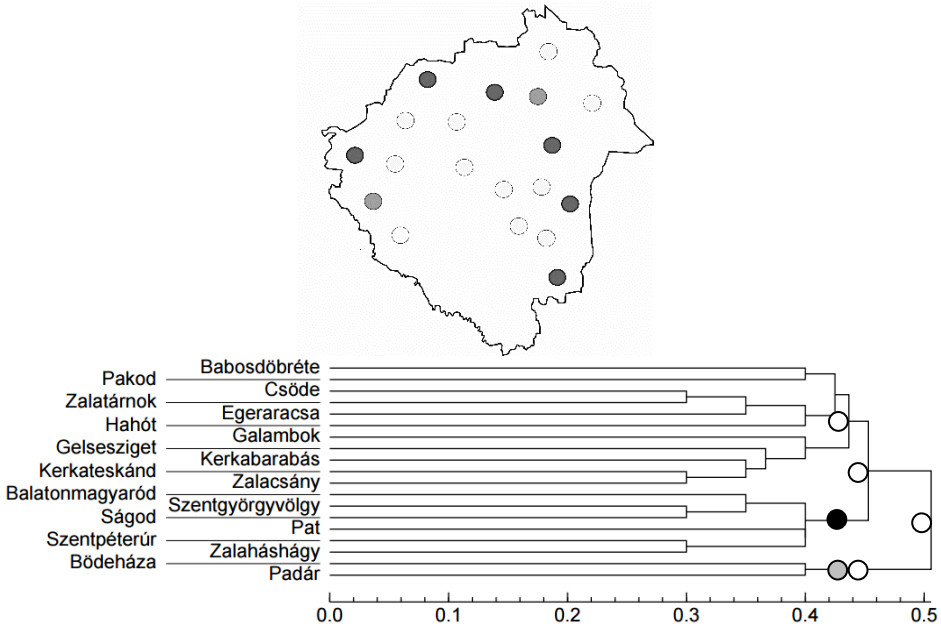
13. ábra: A klaszteranalízis során létrehozott második dendogram

Minthogy az öt település névrendszerét felölelő elemzés végére csupán két csoportunk maradt, ezeket összekötve megkapjuk a vizsgált korpuszra vonatkozó végleges fagráfot (lásd a 14. ábrán).



14. ábra: Zala megye 5 településének végleges dendogramja

A csoportátlag módszer eredményét a 15. ábrán összegzem. Abból adódóan, hogy ez a módszer nem a legkisebb vagy éppen a legnagyobb távolságot használja fel, hanem igyekszik egy teljes, átlagos képet adni a névrendszerről, előnyben részesíthető az egyszerű vagy a teljes láncmódszer használatával szemben.

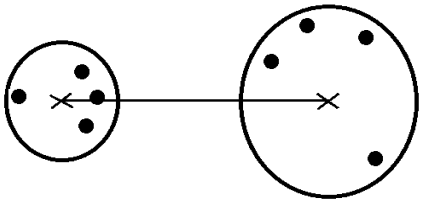


15. ábra: A csoportátlag módszer térképe és dendogramja Zala megye 19 települése alapján

A dendogramból megállapítható, hogy Padár és Bödeháza kiugró értékeket képvisel, adataik ugyanis a többi zalai településtől jobban eltérő csoportot alkotnak (a térképen és a dendogramon szürke körrel jelöltem őket). A fekete körrel ábrázoltak pedig illeszkednek a Zala folyó vonalára.

4.2.4. A klaszteranalízis negyedik eljárása az ún. centroid vagy súlypont módszer. E metódus során igyekszünk a névrendszerekben megfigyelhető szignifikáns különbségek alapján összevetni az elemeket, s ez lesz a centroid. Az analízis

során a centroidok közti távolságok szerint alakíthatók ki az egyes csoportok (lásd TAN–STEINBACH–KUMAR 2006: 517).



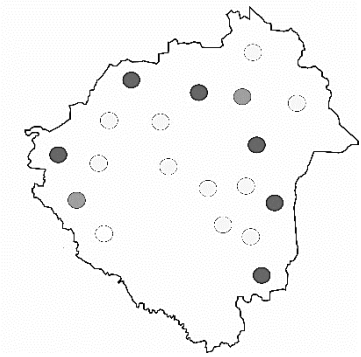
16. ábra: A centroid módszer sematikus ábrája

Ez a módszer alkalmazható nyelvészeti, névtani kutatásokra is, ugyanakkor azért nem a legmegfelelőbb az ilyen irányú vizsgálatokhoz, mert első lépésként az összes adatunkat átlagolja (ez az átlag lesz az egyes csoportok centroidja), majd ezeket az átlagértékeket hasonlítja össze. Ennél az eljárásnál a csoportátlag módszer lényegesen érzékenyebb. Emellett további hátránya, hogy ha egy klaszteren belül erősen eltér az elemszám (azaz egy klaszterbe különböző számú települések kerülnek), az nagyban befolyásolhatja az eredmény valósságát. Az a körülmény azonban, hogy ez az eljárás tulajdonképpen ugyanazt az eredményt nyújtja, mint a csoportátlag módszer (lásd ehhez a 17. ábrát), csupán megerősíthet minket a csoportátlag módszer releváns voltáról.

4.2.5. Végezetül a klaszteranalízis során igen gyakran alkalmazott eljárásról, a variancia módszerről, vagy másképpen a Ward módszerről kell még szólnunk. Ez a módszer nagyon hasonlít a csoportátlag módszerhez, tulajdonképpen a pontok közötti távolságok vagy hasonlóságok négyzetével kell számolnunk. Ebben az esetben két klaszter hasonlósága a klaszteregyesülések során fellépő eltérésnégyzet növekedésén, vagy hasonlóságnégyzet csökkenésén alapszik.[4] Ez az eljárás nem érzékeny a hibás vagy kiugró adatokra. Korlátozottan használható azonban akkor, ha a csoportok közti különbség erősen eltér, de akkor is, ha a kialakuló klaszterek mérete nagyon különbözik (vö. TAN–STEINBACH–KUMAR 2006: 523).

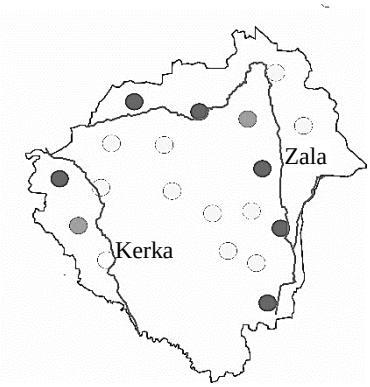
A variancia módszert a névrendszerekre véleményem szerint a csoportátlag módszerhez való nagyfokú hasonlósága miatt lehet alkalmazni, azonban megbízhatóságát csökkenti az, hogy a különböző névrendszerek akár nagyon eltérő méretű csoportokat is alkothatnak, márpedig ez a módszer nem a legmegfelelőbb az ilyen típusú klaszterek összevetésére.

[4] Ez csupán attól függ, hogy a kialakított mátrix a hasonlósági vagy a különbségi fokot mutatja. Két település tehát lehet 0,7 mértékben hasonló, azaz 0,3 mértékben különböző. A mátrixok kialakítása mindkét módon történhet, a klaszteranalízis azonban — mint azt korábban is jeleztem — döntően különbségmátrixokon alapszik.



17. ábra: A csoportátlag módszer, a centroid, valamint a variancia módszer eredményeit szemléltető térképlapok

A Zala megyei települések névrendszereire vonatkozó analízis eredményei azt mutatják, hogy a csoportátlag módszer, a centroid és a variancia módszer is azonos eredményeket rajzol ki a névföldrajzi tájakról (lásd ehhez a 17. ábrát), s dendogramjuk is azonos (megegyezik a 15. ábrán bemutatott dendogrammal). (A fekete körök az egymással azonos csoportba tartozó településeket jelölik, a szürkék pedig a többitől jelentősen eltérőeket.) Az a körülmény, hogy három, valamelyest eltérő klaszterező módszer is végső soron rendkívül hasonló eredményt hozott, arra enged következtetni, hogy a fenti ábrán bemutatott helyzet mögött valós névföldrajzi eltérések sejthetők. A klaszteranalízis alapján azonos csoportba tartozó települések a Zala és talán a Kerka folyók vonalán helyezkednek el (lásd 18. ábra), két település (Bödeháza és Padár) névrendszere pedig a többitől nagyban különbözik.



18. ábra: A klaszteranalízis során kapott eredményt szemléltető térkép a Zala és a Kerka folyókkal.

A 19 Zala megyei település névrendszerének klaszteranalízise azt mutatja, hogy a Zala és talán a Kerka folyók völgye valamiképpen összefüggő névföldrajzi

területet alkothat. Ezt a megfigyelésünket később azáltal pontosíthatjuk, illetőleg árnyalhatjuk, hogy a megye reprezentatív mintavétellel meghatározott településeinek helynévrendszerét bevonjuk a vizsgálatba, de akár a megye összes településének ilyen jellegű analízise is kirajzolhat eredményeket. Ebben az írásomban én azonban csupán arra vállalkoztam, hogy a klaszteranalízis lehetséges eljárásait bemutassam, s ezek között a célszerűbb módszerekre rávilágítsak. Az pedig, hogy ennek során konkrét névföldrajzi területek elkülönítésének a lehetőségét is felvillanthattam, csak további hozadéka lehet írásomnak.

Irodalom

ANDERBERG, MICHAEL R. 1973. *Cluster Analysis for Applications*. New York, Academic Press.

BÁBA BARBARA 2013. A *vejsze* lexéma története és korai ómagyar kori előfordulásai. *Helynévtörténeti Tanulmányok 9:* 43–56.

BÁRTH M. JÁNOS 2010. Helynevek vagy körülírások? *Helynévtörténeti Tanulmányok 5:* 209–221.

BOLLA MARIANN–KRÁMLI ANDRÁS 2012. *Statisztikai következtetések elmélete*. Budapest, Typotex Kiadó.

BRAY, J. ROGER–CURTIS JOHN. T. 1957. An Ordination of the Upland Forest Communities of Southern Wisconsin. *Ecological Monographs 27:* 325–349.

DITRÓI ESZTER 2010. Névrendszerek modellalapú vizsgálata. *Helynévtörténeti Tanulmányok 5:* 155–168.

DITRÓI ESZTER 2011. Egy lehetséges módszer a helynevek területi különbségeinek igazolására. *Helynévtörténeti Tanulmányok 6:* 151–161.

DITRÓI ESZTER 2012. Helynévrendszerek területi differenciáltsága. *Helynévtörténeti Tanulmányok 7:* 29–39.

DITRÓI ESZTER 2013. Nyelvi érintkezések hatása a helynévmintákra. Vendvidéki esettanulmány. *Helynévtörténeti Tanulmányok 9:* 89–100.

DITRÓI ESZTER 2015. A helynévrendszerek területi differenciáltságának statisztikai alapú megközelítése. In: É. KISS KATALIN–HEGEDŰS ATTILA–PINTÉR LILLA szerk., *Nyelvelmélet és dialektológia 3*. Budapest–Piliscsaba, Szent István Társulat. 58–80.

EVERITT, BRIAN 1974. *Cluster Analysis*. London, Heinemann Educational for the Social Science Research Council.

HAJDÚ MIHÁLY 1991. A magyar névtudomány a nyelvjárástörténeti kutatás szolgálatában. In: KISS JENŐ–SZŰTS LÁSZLÓ szerk., *Tanulmányok a magyar nyelvtudomány történetének témaköréből*. Budapest, Akadémia Kiadó. 250–254.

HAJDÚ MIHÁLY 1999. Névutók a helynevekben. *Magyar Nyelvjárások 37:* 187–192.

HOFFMANN ISTVÁN 1993. *Helynevek nyelvi elemzése*. Debrecen, A Debreceni Kossuth Lajos Tudományegyetem Magyar Nyelvtudományi Intézetének Kiadványai 61.

JARDINE, NICHOLAS–SIBSON, ROBIN 1971. *Mathematical Taxonomy*. New York, Wiley.

KÁLMÁN BÉLA 1967. Helynévkutatás és szóföldrajz. *Nyelvtudományi Értekezések 58*: 344–350.

MNyA. = DEME LÁSZLÓ–IMRE SAMU, *A magyar nyelvjárások atlasza 1–6*. Budapest, Akadémia Kiadó, 1968–1977.

PÁRNICZKY GÁBOR 1976. *A statisztikai informatika alapjai*. Budapest, Statisztikai Kiadó Vállalat.

ROBINS, ROBERT HENRY 1999. *A nyelvészet rövid története*. Budapest, Osiris Kiadó–Tinta Kiadó.

TAN, PANG-NING–STEINBACH, MICHAEL–KUMAR, VIPIN 2006. *Inroduction to data mining*. Boston, Pearson Education.

TÓTH VALÉRIA 1998. Ómagyar helyneveink és a névföldrajz. *Magyar Nyelvjárások 35:* 121–134.

TÓTH VALÉRIA 2001. *Névrendszertani vizsgálatok a korai ómagyar korban (Abaúj és Bars vármegye)*. A Magyar Névarchívum Kiadványai 6. Debrecen, Debreceni Egyetem Magyar Nyelvtudományi Tanszék.

TÓTH VALÉRIA 2002. A helynévmodellek nyelvföldrajzi vizsgálata a korai ómagyar korban. In: HOFFMANN ISTVÁN–JUHÁSZ DEZSŐ–PÉNTEK JÁNOS szerk., *Hungarológia és dimenzionális nyelvszemlélet.* Debrecen–Jyväskylä, Debreceni Egyetem, Magyar Nyelvtudományi Tanszék. 127–138.

VARGHA FRUZSINA SÁRA 2010. A dialektometria alkalmazása és történeti helynevek nyelvföldrajzi vizsgálata a Székelyföldön. *Helynévtörténeti Tanulmányok 5:* 223–233.

VARGHA FRUZSINA SÁRA–BODÓ CSANÁD–VÉKÁS DOMOKOS 2012. Classifications of Hungarian dialects in Moldavia. In: LEHEL, PETI–VILMOS, TÁNCZOS szerk., *Language Use, Attitudes, Strategies: linguistic identity and ethnicity in the Moldavian Csángó villages*. Cluj-Napoca, The Romanian Institute for Research on Minorities. 51–69.

VARGHA FRUZSINA SÁRA–VÉKÁS DOMOKOS 2012. Lokalizálható nyelvtörténeti adatok informatizálása és térképezése. *Erdélyi Múzeum 74:* 160–165.