Számítógépes dialektológia

Mi a számítógépes dialektológia, és miért fog hamarosan "megszűnni", éppen akkor, amikor az informatikailag is megalapozott nyelvészeti módszerek révén és mentén alakulhatnak ki széleskörű együttműködések, amikor az új technológiákra építve indulhatnak nagyszabású dialektológiai vállalkozások? A következőkben erre a kérdésre keresem a választ, áttekintve a számítógépes dialektológia hazai és Kárpát-medencén túli legjellemzőbb állomásait, céljait, eredményeit.

1. Mitől számítógépes a dialektológia? – A számítógépes dialektológia felfogható egy nyelvészeti módszerként, a feladatok megközelítési és hatékony megoldási módjaként, amely a terepmunka előkészítésétől az adatok informatizálásán és feldolgozásán keresztül az elemzésig segíti a kutatást megfelelő nyelvészeti technológiák tervezésével, fejlesztésével és alkalmazásával. Elsődleges célja informatizált, azaz megfelelő digitális formába hozott, s így sokoldalúan hasznosítható és újrahasznosítható nyelvi adatok létrehozása.

A számítógépes dialektológia lényegét úgy is megérthetjük, ha megnézzük, mi nem az. A puszta számítógép-használat nem jelent számítógépes nyelvészetet: hiába vannak korszerűnek számító informatikai eszközeink, ha nem informatizált adattárak létrehozása a célunk. Lássunk egy példát: Wolfgang Viereck előszeretettel használja a számítógépes dialektológia (computational dialectology) kifejezést, ha azonban 1997-es hozzáállását tekintjük, a tipikus számítógép-használó dialektológus, és nem a számítógépes dialektológus képe jelenik meg előttünk. Viereck ugyanis végtermékként mindössze papíratlaszt, Anglia nyelvatlaszát generálta számítógéppel. Az atlasz készítésének folyamatát bemutató tanulmányában (VIERECK 1997) a kilencvenes évek elején még viszonylag újnak számító lézer-nyomtatók egyik tulajdonságát, az úgynevezett postscript képességet említi (nem kevesebb mint tizenhárom alkalommal), mint a számítógépes dialektológiát előrelendíteni hivatott vívmányt. Eközben valóban aggasztó nehézségekről számol be meglepő őszinteséggel, és számunkra igen tanulságosan: ahogy az évek telnek, az informatikai platformok változnak, és a különböző platformok közötti adatmozgatás az információ érthetetlen deformálódásával járt esetükben. Ráadásul az adatrögzítők változatos megoldásokkal rögzítették az azonos jelenségeket, így a konzisztencia sérült. Mindezek miatt többször elölről kellett kezdeni a műveleteket, jelentős anyagi és időveszteséggel. A történet sűrítve mutatja a nyelvészeti célú számítógép-használat kockázatait, miközben világosan kiderül, hogy az ismertetett projektum sem a célkitűzéseit, sem a módszereit tekintve nem minősül számítógépes dialektológiai vállalkozásnak (hiszen akkor a fő célja éppen az lenne, hogy az informatizált adatok a platformok között egyszerűen hordozhatók legyenek, és megfelelő technológiákkal és eljárásokkal vissza lehessen szorítani az adatrögzítési anomáliákat.) Noha kiváló dialektológus és postscript-szakértő informatikus is dolgozott az angol nyelvatlaszon, a csapat az informatikai technológiákat dicsőítve (ugyanakkor saját magukat ugyanezen technológiák tehetetlen áldozatának is feltüntetve) küszködött hosszú éveken keresztül egy kartográfiai feladattal.

2. Mozzanatok a számítógépes dialektológia történetéből. – Noha voltak már korábbi kezdeményezések is lejegyzett nyelvjárási adatok adatbázisba rendezésére, a korszerű értelemben vett számítógépes geolingvisztika születése elválaszthatatlan a kaliforniai UCLA-től: a nyolcvanas évek végén, amikor megjelentek az első multimédiás számítógépek (a legendás Apple Macintosh-ok), itt született meg a terepen gyűjtött és digitálisan tárolt olyan nyelvi adat fogalma, amely földrajzi ponthoz, területhez köthető, fonetikusan lejegyzett, ugyanakkor

hangzó formájában is elérhető, akár akusztikailag is elemezhető, méghozzá magán a számítógépen futó programmal, nem külső eszközzel (LADEFOGED – MADDIESON 1996). A fejlesztések vezetője, egyben a legfőbb fejlesztő Peter Ladefoged, a Nemzetközi Fonetikai Társaság akkori elnöke volt, személyes példája és iránymutatása a későbbiekben a Bihalbocs tervezésére is hatott.

Mindössze néhány évvel a Los Angeles-i kezdetek után Balogh Lajos és munkatársai, abból kiindulva, hogy A magyar nyelvjárások atlasza papír formájában holt anyag, megtervezték és elkezdték a nagyatlasz informatizálását (BALOGH – KISS 1992). Módszerük elméletileg teljes értékű informatizálást tesz lehetővé (eltekintve az informatikai platformok változásai kapcsán fellépő kompatibilitási problémáktól). Kezdeményezésük jól példázza, hogy eltökéltséggel és szorgalommal akkor is neki lehet látni fontos számítógépes dialektológiai részfeladatoknak, ha nincsenek még beszerezhető technológiák, és a saját fejlesztések feltételei is hiányoznak. A sokoldalú számítógépes dialektológiai technológiák fejlesztése ugyanis roppant erőforrásokat igénylő feladat, ilyen erőforrások az eddigi tapasztalatok szerint pályázati keretekből nem teremthetők elő.

Balogh Lajos kezdeményezése nem bizonyult eléggé vonzónak a lehetséges követők számára: újabb nyelvjárási adattárak megfelelő számítógépes rögzítésére nem került sor, a nagyatlasz informatizált változata is torzóban maradt. A korszerű értelemben vett számítógépes dialektológia hazai indulásához még szükség volt egyrészt a kockázatokkal is számot vető, bátor tudománypolitikai döntésre és elkötelezettségre (ez Kiss Jenő nevéhez fűződik), másrészt a szükséges nyelvészeti technológiák megtervezésére és kifejlesztésére. A kettő egymást feltételezi: nem lehet tudománypolitikai döntést hozni, ha a sikeres megvalósítás lehetősége nem körvonalazódik, mint ahogy nem érdemes rendkívül erőforrásigényes tervezésbe és fejlesztésekbe kezdeni, ha nem körvonalazódik kutatók fogadókész közege. A Bihalbocs néven ismertté vált, 1996-ban megindított fejlesztések sokáig kizárólagos teherviselőjeként arra volt szükségem, hogy lássam, van fogékonyság, van általános elkötelezettség, és van megfelelő nyitottság, további igények is felkelthetők.

A Bihalbocsban a kezdetektől alapvető fontosságú a hanganyag és lejegyzés kapcsolatának biztosítása, másrészt az olyan térinformatikai eljárások megteremtése, amelyek a papíratlaszoknál összehasonlíthatatlanul hatékonyabban teszik lehetővé nyelvjárási adattárak kutatását, elemzését. Az elemzésnek azonban nyilvánvaló előfeltétele, hogy legyenek feldolgozható adataink. A dialektológiai fejlesztéseknek tehát legelőbb az adat-informatizálás feltételeit kellett megteremteniük. Így folytatódhatott Balogh Lajos, majd Bodó Csanád irányításával a nagyatlasz térképlapjainak informatizálása is.

A továbbiakban lássunk néhány kiemelkedő példát arra, milyen lehetőségeket látnak mások a számítógép geolingvisztikai alkalmazásában.

Egy klasszikus filológiai képzettségű és matematikai fogékonyságú tipikus számítógépes dialektológus, Hans Goebl salzburgi romanista 1984-es disszertációjában számos dialektometriai eljárást dolgozott ki, amelyek nagy részét 2000 után egy munkatársa számítógépes programba építette. A salzburgi romanisztikai műhely és dialektometriai (kvantitatív nyelvföldrajzi) iskola az egyes nyelvjárások közötti hasonlóság (vagy eltérés) mérését tűzi ki célul a hagyományos atlaszlapokból elvont ún. munkatérképek kvantitatív szintézisével (GOEBL 2006). Goebl és munkatársai, miként mások is, abból indulnak ki, hogy az izoglosszák klasszikus eszköze túlzottan önkényes; objektívebb megoldásokra van szükség. A munkatérképeken a kutatópontokhoz már nem nyelvi adatok, hanem számok tartoznak, amelyek a vonatkozó adatoknak a vizsgált jelenség szerinti csoportosítását tükrözik. A kutatópontok közötti szám-megfelelések összesítése egy hasonlósági mátrix, amely alapján az erre a célra kifejlesztett program térképre vetíti egy kiválasztott kutatópont más kutatópontok nyelvjárásához való, színskálával érzékeltetett számszerű hasonlóságát. Megjegyzendő, hogy a salzburgi módszerrel rokon eljárással a már informatizált és megfelelően csoportosított adatokból

(pl. Király Lajos Somogy-zalai nyelvatlaszának már informatizált adattárából) interaktív hasonlósági térképek hozhatók létre.

Más dialektometriai elképzelésekhez (pl. HEERINGA 2004) nem az adatok csoportosítása szolgál alapul, hanem a betűláncnak felfogott, fonetikusan lejegyzett nyelvi adatok páronkénti összevetése és a távolság számszerűsítése egy matematikai eljárással ("string edit distance" módszerek). A legismertebb a Levenshtein-algoritmus, magyar nyelvjárási anyagon való alkalmazásának célszerűségére Kiss Jenő hívta fel a figyelmemet. A módszer előnye, hogy az adatok nem elég pontos és nem mindig objektív csoportosításának munka- és időigényes feladata mellőzhető, a nyelvjárások közötti affinitás mértékének számszerűsítését az algoritmus nyelvészeti szempontoktól meghatározott informatikai megvalósítására bízzuk. (Megjegyzendő, hogy Goebl nem fordított figyelmet a nyelvi adatok informatizálására, ezért munkatérképei, ahol csak a klasszifikáció numerikus eredményei kapcsolódnak a kutatópontokhoz, Levenshtein-algoritmussal nem vizsgálhatók.)

LABOV és munkatársai Észak-Amerika nyelvváltozatait vizsgáló atlaszát (2006) a fonetikai nézőpont dominanciája jellemzi. Magánhangzó-formánsok térképre vetítésével olyan célt valósít meg, amely igényként A magyar nyelvjárások atlasza munkálatai során is jelentkezett, tervezték ugyanis, hogy eszközfonetikai vizsgálatok eredményeit is térképezik (BENKŐ 1975: 132). Labovék atlaszának másik, számunkra is fontos tanulsága, hogy a nyelvföldrajz és az általános nyelvészet közötti kapcsolódások újbóli megteremtését elsődleges célként kezeli. Az általános nyelvészet, fonetika és fonológia nehezen fejlődhetne a nyelvjárási jelenségek gazdag tárházának hasznosítása nélkül.

Nem csak Észak-Amerikában, hanem Európában is egyre inkább meghatározóvá válik a fonetikai központú megközelítés. Ennek megfelelően a hangzó, akusztikailag elemezhető nyelvjárási (területhez kötött) adatok az elsődlegesek, abban az értelemben is, hogy a hangzó anyag van meg előbb, és annak digitális formájából jön létre az átírt adat, és a tekintetben is, hogy elemzéskor a hanganyaghoz nyúlunk vissza, nem elégszünk meg a lejegyzett forma értékelésével. Hanganyag és lejegyzés összekapcsolása tehát megkerülhetetlen feladattá vált a dialektológiai kutatásokban. CorPho néven 2006-ban európai együttműködés indult a hangtani kutatások igényeinek is megfelelő, jellemzően nyelvjárási hangfelvételeket tartalmazó beszéltnyelvi korpuszok fejlesztésének összehangolására. A multilingvális nagy adattárakhoz ugyanis a kisebb projektek közötti együttműködés szükséges, ehhez pedig az adatok kódolásának és a metaadatoknak a szabványosítása vinne közelebb. A kérdéskör összetettsége miatt azonban a szabvány kialakítása nem megy egyik napról a másikra, a divatos varázsszavak (pl. XML, Text Encoding Initiative) nem jelentenek megoldást.

3. Nyelvészeti technológiák alkalmazása a magyar dialektológiában. – Lássuk röviden, melyek az aktuális főbb irányai a magyar számítógépes dialektológiának. Elsőként a papíratlaszok informatizálását említhetjük, ami már sok éve ütemesen halad. A papíron kiadott tájszótárak informatizálása a KISS JENŐtől (2002) felvázolt módon a közeljövőben kezdődhet, az erőforrások függvényében. Eltekintve a részletes bemutatástól, azt emelném ki, hogy e munka- és időigényes feladatok is szükségesek a dialektológia új fejlődési pályára állításához: ma már nem a papírra nyomtatott, hanem az informatizált (tehát sokoldalúan felhasználható, többek között papírra is nyomtatható) adatot érdemes alapértelmezettnek tekinteni.

A lényegesen összetettebb, hanganyagokat is feldolgozó projektumok közül kettőt nevezek meg. Egyik a MNyA. hangfelvételeinek részleges feldolgozása (lásd bővebben VARGHA, e kötetben), a másik pedig egy moldvai multimédiás atlasz létrehozását tűzi célul, egyben követéses vizsgálat, és a jelenleg folyó legösszetettebb számítógépes dialektológiai vállalkozás, nemzetközi összehasonlításban is rendkívül újszerű. Ez utóbbinak két legfontosabb jellemzője egyrészt a korszerű szociolingvisztikai szempontrendszer erőteljes jelenléte, másrészt az új nyelvészeti technológiákra épülő adatgyűjtés és feldolgozás (lásd bővebben BODÓ, e kötetben).

Az új nyelvészeti technológiák segítik elő a szomszédos tudományterületek közötti együttműködést is, például az informatizált nyelvjárási és a lokalizálható nyelvtörténeti (névtani) adatok integrálásával (VARGHA 2007).

4. Miért van szükség dialektológiai szoftverek fejlesztésére? – Az egyre összetettebb szaktudományos feladatok megoldásához speciális szoftverek használata szükséges, ám nincs a polcról levehető, a boltban megvehető megoldás. A kereskedelmi forgalomban beszerezhető szoftverek ugyanis (szövegszerkesztők, hangszerkesztők, adatbázis-kezelők, grafikus programok, térinformatikai alkalmazások) csupán a kutatói igények töredékének kiszolgálására lennének alkalmasak.

Nyelvjárási adatainkat valamilyen adattárban, adatbázisban tároljuk, és az adatokból készített egyszerűbb vagy bonyolultabb kimutatások eredményét térképre kívánjuk vetíteni. A szokásos adatbázis-kezelők nincsenek felkészítve a térképezési feladatra, ezért kevéssé hasznosak. A Bihalboccsal azonban akár különböző adattárakból származó adatok integrálása is lehetséges egy térképen, hiszen a különböző adattárak kutatópontjai – a földrajzi koordináták egyetemessége révén – ugyanarra a térképre is vetíthetők. Az adatintegrálás lehetősége az informatizálás alapvető haszna.

A tudományterület követelménye az adatok hangzó formájának azonnali elérése. Az általános célú adatbázis-kezelők azonban nem alkalmasak hang és szöveges adat egymáshoz rendelésére, ami – legalábbis az igényeknek megfelelő módon – túl bonyolult feladat, így ez ránk, a tudományterület művelőire hárul.

Folyamatos nyelvjárási szövegek esetén a lejegyzés és a hangzó változat összekapcsolása általában a következőképpen történik: a hangfájl áll a középpontban, hiszen az az eleme egy adattárnak, és így a hangfájlnak van lejegyzése időzítési markerekkel, amelyek időben lineárisan követhetik egymást. Egy ilyen gyakorlathoz különböző szoftverek is találhatók, azonban az ilyen egyszerűsített, rugalmatlan megoldások a kutatási lehetőségeket korlátozzák. Ezért a Bihalbocs megfordította a perspektívát: nem a hangfájlnak (ami nem nyelvészeti kategória), hanem a lejegyzett interjúnak és adatnak van központi szerepe, a szövegben elhelyezett időzítési markerek tetszőleges sorrendben tetszőleges hangfájlok tetszőleges időpillanataira utalnak.

A fonetikusan lejegyzett adatok informatizálása történhet ugyan szokásos szövegszerkesztővel is, ám az erre a feladatra optimalizált dialektológiai szoftver jelentősen növeli az adatrögzítés hatékonyságát. Ennél fontosabb szempont, hogy a fonetikai szimbólumok értelmezésén alapuló adatkezelés és konverzió csakis speciálisan felkészített számítógépes programmal lehetséges. A nyelvjárási lejegyzés-szerkesztés főbb elvei a Bihalbocsban felsorolásszerűen a következők: nyelvészeti szempontú analitikus kódolás, a szerkesztéskor grafikusan összerakható jel egységének megőrzése, a mellékjelek grafikai variánsai és kötött sorrendjük, összetett jel esetén a mellékjeleknek az utolsó elemhez történő hozzárendelése. A Bihalbocs ezekre az elvekre épülve biztosít szilárd alapot nyelvjárási adatok sokoldalú használatához, s az adatok szükség szerinti továbbkonvertálásához. Ha egy másik rendszert használunk, nem biztos, hogy informatizált adataink lesznek. Ha ugyan egy másik rendszerben, de világos struktúrákban és egyértelmű kódolással végzünk adatrögzítést, az is jó megoldás lehet, hiszen a későbbiekben szükségessé váló konverzió vélhetőleg megoldható lesz.

5. Összefoglalás. – A számítógépes dialektológia új megközelítést kínál, de az új célok mellett, azokkal szoros összefüggésben, a már korábban kigondolt, de esetleg nehezen elérhetőnek tűnő feladatok megvalósítását kívánja megkönnyíteni, így a korábbi időszakok eredményeit tudja és kívánja valorizálni (például azzal, hogy létrehozzuk egy papíratlasz összehasonlíthatatlanul sokrétűbben felhasználható informatizált változatát, vagy hanganyagok pusztulásra ítélt tárát alakítjuk át egy feldolgozott, kereshető, akár akusztikailag is vizsgálható beszéltnyelvi korpusszá). Törekvéseink értelmében az új nem áll szemben a régivel, inkább felkarolja azt.

Mivel az új módszerek és technológiák már ma is befolyásolják, sok szempontból pedig meg is szabják, hogy milyen dialektológiai feladatokat és milyen sorrendben tűzhetünk ki, és milyen módon végezhetünk el, ezért a számítógépes dialektológia felfogható a dialektológiai kutatások fősodraként.

A nyelvészeti technológiák fejlesztői leginkább arra törekszenek, hogy minél eredményesebben és hatékonyabban oldják meg a felmerülő nyelvészeti feladatokat, és ezzel új kutatási lehetőségeket tárjanak fel, nem csak maguknak, hanem mindenkinek, aki élni kíván ezekkel a lehetőségekkel (vállalva a szükséges idő- és energiaráfordítást is, amihez képest eltörpül a hangfelvevő, a számítógép és az adattárolók költsége). Természetesen cél továbbá, hogy olyan informatikai eszközöket fejlesszünk (vagy ha már ilyen van, megtaláljuk, s használatba vegyük), amelyek minél többek számára teszik lehetővé, hogy geolingvisztikai projektekben részt vegyenek, hogy adataikat sokoldalúan hasznosítható formába hozzák.

A számítógépes dialektológia ma még önálló fejezet a dialektológia tankönyvben (KISS 2001), hiszen Kiss Jenő így is alá akarta húzni e megközelítés fontosságát és önállóságát; a következő tankönyvben azonban már minden bizonnyal nem lesz ilyen fejezet. Várakozásaim szerint a számítógépes módszerek éveken belül standardizálódnak, és egyre szélesebb körben elterjednek, használatuk nem jelent majd megkülönböztető jegyet. A számítógépes dialektológusok éppen azon dolgoznak, hogy a számítógépes dialektológia hamarosan megszűnjön, hogy éveken belül feleslegessé váljon a számítógépes jelző.

Hivatkozott irodalom

BALOGH LAJOS – KISS GÁBOR 1992. A magyar nyelvjárások atlaszának számítógépes feldolgozása. In KONTRA szerk. 1992: 5–17.

BENKŐ LORÁND 1975. A magyar nyelvjárások atlaszának hangjelölési rendszere és gyakorlata. In: DEME – IMRE szerk. 1975: 123–165.

BODÓ CSANÁD. Követéses geolingvisztikai vizsgálat Moldvában. E kötetben.

DEME LÁSZLÓ – IMRE SAMU szerk. 1975. A magyar nyelvjárások atlaszának elméleti-módszertani kérdései. Akadémiai Kiadó, Bp.

GOEBL, HANS 2006. Recent Advances in Salzburg Dialectometry. Literary and Linguistic Computing 21: 411–435.

HEERINGA, WILBERT JAN 2004. Measuring Dialect Pronunciation Differences using Levenshtein Distance. Groningen Dissertations in Linguistics 46.

KISS JENŐ szerk. 2001. Magyar dialektológia. Osiris Kiadó, Bp.

KISS JENŐ A számítógépes dialektológia. In: KISS szerk. 2001: 141–144.

KISS JENŐ 2002. Tájszótárírás és tájszótárak. Magyar Nyelvőr 126: 391–415.

KONTRA MIKLÓS szerk. Társadalmi és területi változatok a magyar nyelvben. MTA Nyelvtudományi Intézet, Bp.

LABOV, WILLIAM – ASH, SHARON – BOBERG, CHARLES 2006. The Atlas of North American English: Phonetics, phonology and sound change. Mouton / de Gruyter, Berlin.

LADEFOGED, PETER – MADDIESON, IAN 1996. Sounds of the World's languages. Blackwell, Oxford, UK / Cambridge, MA.

VARGHA FRUZSINA SÁRA. Nyelvi változók A magyar nyelvjárások atlasza hangfelvételeiben. E kötetben.

VARGHA FRUZSINA SÁRA 2007. Nyelvjárási és helynévtörténeti anyagok számítógépes feldolgozása. Kézirat. Megjelenik a Kontextus–Filológia–Kultúra II. konferenciakötetben. Besztercebánya.

VIERECK 1997. The Computer Developed Linguistic Atlas of England, volumes 1 (1991) and 2 (1997): Dialectological, computational and interpretative aspects. ICAME Journal 21: 79-90.

VÉKÁS DOMOKOS