Audio kazettáról mesterséges intelligencián alapuló algoritmusba

Veszélyben lévő kutatási adatok megóvása – beszámoló egy pilot projektről és az eredmények további sorsáról

Egyed-Gergely Júlia[1]
ORCID: 0000-0003-1905-0027

Jakab Miklós[1]

Meiszterics Enikő[1]

[1] ELKH Társadalomtudományi Kutatóközpont
Kutatási Dokumentációs Központ

Absztrakt

A Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központjában (TK KDK) a folyó kutatások adatkezelési igényeinek és az adatok másodfelhasználóinak kiszolgálása mellett intenzív archiválási munka is zajlik, hiszen a KDK és a keretei között működő 20. Század Hangja Archívum és Kutatóműhely az elmúlt 50 év társadalomtudományos kutatási anyagait gyűjti. A Hungarian Research Data Alliance (HRDA) és a Magyar Tudományos Akadémia Könyvtár és Információs Központ (MTA KIK) által szervezett, az Eötvös Loránd Kutatási Hálózat (ELKH) Titkársága által támogatott adatarchiválási pályázatnak köszönhetően az intézmény munkatársainak lehetősége nyílt a KDK digitális tárházát bővíteni. A projekt keretében két értékes, a digitális korszak előtt készült, a kutatói- és a nagyközönség számára eddig nehezen elérhető társadalomtudományos kutatási anyag vált – a FAIR alapelveket követve – kutathatóvá. Ezzel, a két vizsgálat publikált eredményein túl, most már azok háttéranyagai, az azokban gyűjtött adatok, készített interjúk is megtekinthetővé, újrafelhasználhatóvá, másodelemezhetővé, az újabb kutatási eredményekkel összevethetővé, közkinccsé váltak.

A két archivált anyag ráadásul, immár digitalizált formában, a KDK és a SZTAKI közös MILAB „Computational Archival Science" projektjébe is bekerült. A projekt célja bővíteni a KDK-ban elérhető nagy mennyiségű kutatási dokumentum metaadatkészletét a digitálisan olvasható interjúk gépi tanulás és mesterséges intelligencia segítségével történő, tárgyszavakkal való ellátásával.

Az alábbi tanulmány a HRDA pilot projekt megvalósulását és a TK KDK archívumába bekerült új kutatási anyagok további sorsát mutatja be.

Bevezető

A Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központja két adatrepozitóriumot működtet, amelyek kutatási adatok és dokumentációk gondozásával, elérhetővé tételével foglalkoznak. A KDK-repozitórium a TK négy kutatóintézetének kvalitatív és kvantitatív módszerekkel készült anyagait tárolja (pl. interjúfelvételeket, leiratokat, vezérfonalakat, kérdőíves felmérések kérdőíveit, módszertani leírásokat, adatbázisokat, terepnaplókat, jegyzőkönyveket) különféle formátumokban (pl. szöveg, kép, video). A 20. Század Hangja Archívum és Kutatóműhely az elmúlt hatvan év kvalitatív módszerekkel készült, szociológiatörténetileg is meghatározó kutatásainak anyagait gyűjti írott, hangzó és képes dokumentumok formájában. A két repozitórium anyagait leíró metaadatokat bárki szabadon böngészheti, magukhoz a letétbe helyezett kutatási adatokhoz, dokumentációkhoz az elhelyező döntésének függvényében a KDK repozitóriumában szabadon, regisztrációval vagy egyedi kutatói engedéllyel, a 20. Század Hangja Archívumban regisztrációt követően lehet hozzáférni.

Az archívumok állománya folyamatosan bővül – új és évtizedekkel ezelőtt befejezett kutatások anyagaival egyaránt. Utóbbiakra azért is helyezünk különös hangsúlyt, mert az idő előrehaladtával ezek az anyagok egyre veszélyeztetettebbekké váltak, válnak, folyamatosan veszítenek eredeti minőségükből, ráadásul a technológiai fejlődés miatt a használatukat lehetővé tévő technikai eszközök, berendezések is egyre elérhetetlenebbek, tönkremennek, megsemmisülnek.

A HRDA és az MTA KIK pályázati kiírása kiváló lehetőség volt az ilyen régi anyagok digitalizálásra és közkinccsé tételére. A projekt keretében a KDK két, a múlt század 80-as, 90-es éveiben végzett kutatás audio kazettán lévő interjús hanganyagát tette hozzáférhetővé repozitóriumából – digitalizálás, leiratozás és metaadatokkal való ellátás után.

A két kutatásról

Kovách Imre Kuczi Tiborral közösen folytatott *A helyi társadalom és a mezőgazdaság átalakulása a rendszerváltás idején* című kutatásának témája a termelőszövetkezetek felbomlása a 80-as, 90-es években, valamint a társadalmi-gazdasági változásokra adott válaszlehetőségek, az érintett magánemberek és „háztájizók" megoldási kísérletei. Kovách és Kuczi a vizsgálathoz interjúkat készítettek mezőgazdasági termelőkkel és kistermelőkkel két hullámban és két térségben. Az adatfelvételek első hulláma a 80-as évek közepén, második a 90-es évek elején zajlott, a helyszínek a Dunántúl (Bajna és Epöl), illetve Hajdú-Bihar megye (Hajdúnánás és Hajdúböszörmény) voltak. A kutatás eredményei megjelentek Kovách Imre *A jelenkori magyar vidéki társadalom szerkezeti és hatalmi változásai* című MTA doktori értekezésében (Kovách 2010), valamint a Kovách Imre és Megyesi Boldizsár által közösen jegyzett *A vidék harminc éve. A magyar vidék alakulása az erőforrások, a társadalmi tőke és fejlesztéspolitikai változásainak tükrében* című tanulmányban (Kovách-Megyesi 2018) is. A háttéranyagok eddig nem voltak könnyen elérhetőek az érdeklődők számára, a kutatás során készített interjúk audio kazettán és hagyományos írógéppel készített leiratok formájában várták további sorsukat.

Kovács Éva 1995 és 1998 között szlovákiai vegyes házasságokat vizsgált a *Kulturális csere és etnikai identitásváltozások a vegyesházasságokban a XX. századi kassai népesség példáján* című, Fejős Zoltán által vezetett kutatásban Gyurgyík László, Vasik János, Kádek Kata és Németh Szilvia kutatókkal közösen. A vizsgálatban adatbázisok segítségével, valamint interjúk készítésével és feldolgozásával elemezték a vegyes házasságok demográfiai vonatkozásait 1920-tól kezdődően. Az 1920 és 1991 közötti

időszakra adatbázist hoztak létre a Csehszlovák Statisztikai Hivatal, majd a Szlovák Statisztikai Hivatal által közzétett kiadványok vegyes házasságokra vonatkozó (részben a házasságkötések és válások nemzetiségi bontása, részben a születési adatsorok szülők nemzetiségi bontása szerinti) adataiból, valamint a népszámlálások családi nemzetiségi összetételre vonatkozó adataiból (Gyurgyík 1999, Kovács 2003). Az ezt követő időszakra nem állt rendelkezésre statisztikai adatbázis, a kutatók így részben feltevéseikből indulhattak ki, részben más módszerekben bízhattak. A vizsgálat – egyetemi hallgatókat is bevonva – a házasság és a családi élet témájára összpontosító narratív interjúk készítésével folytatódott. Az interjúk alapján később esettanulmányok készültek a vegyes házasságok életrajzi, társadalmi és családi hátteréről, a családi „magtörténet" narratíváiról, kulturális cseréről, családi traumákról, az identitáspolitika mechanizmusairól és stratégiáiról.[1] A Komáromban készült interjúkat – a kutatás idejének megfelelő technikával – audio kazettán rögzítették.

Kapott anyagok

Kovách Imre vizsgálatából 27 darab 60 perces és 5 darab 90 perces magnókazettán a két hullám adatfelvételeinek hanganyagát és a papíralapú interjúleiratokat kaptuk meg. A projekt indításakor nem lehetett tudni, hogy van-e átfedés a hangzó és az írott anyagok között.

Kovács Éva kutatásából a Komáromban készített interjúk hanganyagát tartalmazó 5 darab 90 perces és 16 darab 60 perces audio kazettát kaptuk meg, valamint ezek mellett papír alapú statisztikai táblákat, rövid elemzést, digitálisan hozzáférhető kutatási jelentést, valamint interjúleiratokat és biográfiákat. Az átvételkor ebben az esetben sem állt rendelkezésre információ arról, hogy van-e, és amennyiben van, milyen mértékű az átfedés a hanganyagok és a leiratok között.

1 Forrás: Társadalomtudományi Kutatóközpont, Kutatási Dokumentációs Központ repozitórium, Fejős-Kovács-Gyurgyík-Vasik-Kádek-Németh kutatási gyűjtemény, absztrakt, https://openarchive.tk.mta.hu/199/

Projektcélok

A projekt célkitűzése a fenti kutatások audio kazettán lévő hanganyagainak és papír alapú dokumentációinak digitális formába történő átírása, a teljes anyag FAIR alapelveknek megfelelő archiválása, valamint a KDK repozitóriumában történő elhelyezése és kutathatóvá tétele volt.

A megvalósítás fázisai

Az audio kazetták vonatkozásában a feladat a két kutatás interjúinak a lehető legjobb kondíciókkal történő hangalapú digitalizálása és leiratozása volt. A hangrögzítés régi 60, illetve 90 perces magnókazettákon történt, a rögzítés mikéntjét, körülményeit, az átvételi jellemzőket nem ismertük. Szakértőkkel egyeztetve a hang digitalizálására és szerkesztésére az Audacity[2] ingyenes és minden platformra elérhető hangszerkesztő programot választottuk. Az Audacity használata egyszerű, menüje, eszköztára átlátható és magyar nyelvű. Az interneten sok, a program használatát a gyakorlatban bemutató videó található. A digitalizált hanganyag leiratozása a Régens Zrt. Alríte[3] programjával történt.

A papíralapú gépelt leiratokat és egyéb papíralapú kutatási dokumentációkat szkennelés után OCR (optikai karakterfelismerés) technológiával, ABBYY FineReader[4] programmal véglegesítettük.

A hanganyag Audacityvel történő digitalizálása

A digitalizálás állandó jelenlétet kívánt, előfordult ugyanis, hogy megszakadt a felvétel, más volt a szalagon, esetleg nem volt rajta semmi. A hangszerkesztő programmal a felvétel közben erősíthettük vagy éppen halkíthattuk a hangot, az optimális hangerősséget digitális kijelzőn követtük. A felvételt WAV tömörítetlen fájlformátumba, majd

2 Audacity hangszerkesztő program, https://www.audacityteam.org/
3 Alríte beszédfelismerő és -leiratozó program, https://alrite.io/ai/hu/
4 ABBYY FineReader karakterfelismerő program, https://pdf.abbyy.com/

– mivel a WAV formátum nagyméretű fájlokban tárolja a digitalizált hanghullámképet, és az Alrite leiratozó programnál szét kellett volna darabolni emiatt az interjúkat – MP3 tömörített fájlba mentettük el. Az Audacityvel való utómunka során a következő korrekciókat végeztük:

- zajcsökkentés (szalagzaj, motorzajok stb.),
- lemezpattogás, torzítások kiszűrése, javítása,
- hangbalesetek (pl. kutyaugatás, mikrofon ütögetése) nyomtalan eltávolítása,
- érthetőség optimalizálása.

A legfontosabb a zajcsökkentés mértékének megfelelő meghatározása volt, ugyanis – egy bizonyos határ után – miközben a zaj csökken, a hasznos jel egyre növekvő mértékű torzulásnak indul. A munkálatok közben sajnos kiderült az is, hogy a felvételek spektrálisan sérültek, ami adódhatott a kazetta minőségéből, a felvétel módjából, eszközéből, idejéből (ami a legvalószínűbb, hiszen a kazetták közel 40 évesek), a tárolásból vagy a többszöri lejátszásból.

A meghallgatás során a felvételeket beazonosítottuk, az egy interjúhoz tartozókat összefűztük, minden interjút metaadatokkal láttunk el. Ezzel lehetővé vált a hangzó interjúk és az átvételkor kapott leiratok összevetése is. A vidék átalakulását vizsgáló kutatás interjúi egy részénél találtunk egyezést, ezek esetében a hanganyagot összekötöttük a papíralapú gépelt leirattal. A komáromi hanganyagok és a kapott digitális leiratok között nem volt átfedés.

A digitalizált interjúk leiratozása az Alrite beszédfelismerő programmal

Az Alrite mesterséges intelligenciára épülő, magyar nyelvre optimalizált beszédfelismerő megoldás, amely napjaink korszerű technikákkal készült hangfelvételeit akár 95%-os pontossággal képes leiratozni. A kapott régi kazetták leiratozása közel sem érte el ezt a technikai szintet, mindössze 10-15%-os pontosságú volt. Szakértő bevonása

céljából felvettük a kapcsolatot a Magyar Rádió egyik hangmérnökével, aki kérésünkre speciális szoftverekkel (MAGIX SEQUOIA[5] 16 és CEDAR Audio[6]) két rövid – ötperces – tesztanyagot készített. A javított hanganyagokon már hallás után azt tapasztaltuk, hogy bár az egyébként is hallható, érthető szövegrészek minősége valóban javult, a nehezen érthető vagy érthetetlen szövegrészek továbbra is nehezen érthetőek, illetve érthetetlenek maradtak, az Alrite használata után pedig továbbra is jelentős mértékben javításra szoruló leiratváltozatokat kaptunk.

A gépi leiratokat ezért csupán támpontként használhattuk, a szöveg nagy részét hallás alapján egészítettük ki, illetve gépeltük be.

A program dolgát nehezítette továbbá, hogy az alanyok sokszor hadartak, tájszólással beszéltek, illetve előfordult, hogy többen (házastársak, szomszédok) egyszerre szólaltak meg, amiket a program nem tudott szétbontani. Gyors beszéd esetén próbáltuk lassítani a felvételt, de ez sem hozott kielégítő eredményt. A tapasztalat az, hogy annál pontosabb a leirat, minél jobb az interjúalany artikulációja – ilyen felvételeknél a régi kazetták esetében is jobb lett a leiratozás minősége, találkoztunk olyan interjúval, amelynél 60%-ban tudta a program hibátlan szöveggé alakítani a hanganyagot.

A hallás alapján való javításnál további nehézségbe is ütköztünk. Amikor az alanyok szakszöveget használtak, szótár segítségét kellett igénybe venni, nemcsak a mesterséges intelligencia nem értette a szöveget, mi sem. Utána kellett nézni az olyan, agronómus interjúalany által használt kifejezéseknek például, mint a *meliorációs munkák*, vagy a *szilázs–szenázs* szópár. Előbbi a talajjavító munkákat, utóbbi a silózásnál a takarmány tartósításának különböző nedvességtartalmát jelenti. Más esetben, amikor a válaszadó az adott város utcaneveit sorolta, a Google-térkép volt a segítségünkre.

5 MAGIX SEQUOIA hangszerkesztő program: https://www.magix.com/int/
6 CEDAR Audio hangszerkesztő program: https://www.cedar-audio.com/

A projektben egy 1 órás régi hanganyag leiratának kiegészítése, javítása jó esetben 3, rossz esetben 6-7 órába telt, az átlag 4 óra volt.

A gépelt leiratok, egyéb kutatási dokumentációk digitalizálása OCR technológiával, ABBYY FineReader programmal

A kapott gépelt leiratokat és egyéb kutatási dokumentációkat OCR technológiával, az ABBYY FineReader program segítségével digitalizáltuk. A szövegek viszonylag jó minőségűek és „tiszták" (kézzel írt bejegyzésektől mentesek) voltak, így az OCR technológiával készített digitális változat esetében a kapott szöveg pontossága – az Alrite mai technológiával felvett interjúk leiratozásának pontosságához hasonlóan – 90% feletti volt. Ez nagy könnyebbséget jelentett azoknál a hangzó interjúknál, amelyeknél rendelkezésünkre állt a leirat gépelt változata is.

Megmentett kutatási anyagok a jelen és a jövő számára

A fent bemutatott technológiák és technikák alkalmazásával a két kiválasztott vizsgálat anyagainak digitalizálása megvalósult, ezzel azok megmenekültek az elkallódás és az olvashatatlanná, hallgathatatlanná válás veszélyétől. Az anyagok gyűjteményekbe rendezve, DOI-val ellátva bekerültek a KDK repozitóriumába, ahonnan a kutatásokat és kutatási adatokat leíró metaadatok, az archiválás módjának leírása, a kutatásokhoz kapcsolódó publikációk, illetve a statisztikai táblák és a módszertani leírások minden érdeklődő számára korlátozások nélkül hozzáférhetőek.

Kovách Imre gyűjteménye a https://openarchive.tk.mta.hu/496/ címen,[7] Kovács Éva gyűjteménye pedig a https://openarchive.tk.mta.hu/199/ címen[8] tekinthető meg.

7 KOVÁCH Imre, KUCZI Tibor: A helyi társadalom és mezőgazdaság átalakulása a rendszerváltás idején. [Kutatási gyűjtemény], https://www.doi.org/10.17203/KDK496.

8 FEJŐS Zoltán, KOVÁCS Éva, GYURGYÍK László, VASIK János, KÁDEK Kata, NÉMETH Szilvia: Kulturális csere és etnikai identitásváltozások a vegyesházasságokban a XX. századi kassai népesség példáján. [Kutatási gyűjtemény], https://www.doi.org/10.17203/KDK199.

Az interjúk leiratait csak a teljes nevek szintjén anonimizáltuk, meglátásunk szerint a települések, városrészek, foglalkozások anonimizálása olyan információvesztést okozna, amely ennyi idővel az adatfelvételek után már nem indokolt. Részben emiatt, részben pedig az interjúkban felmerülő szenzitív témák (pl. a településeken élő szlovák vagy zsidó közösséggel való viszony leírása) miatt magukat az interjúkat nem tettük mindenki számára olvashatóvá, azok kutatói hozzáféréssel érhetőek el. Ez azt jelenti, hogy a Társadalomtudományi Kutatóközpont kutatói korlátozás nélkül, szabadon férnek hozzá az anyagokhoz, külsős kutatók, egyetemi hallgatók pedig igényük jelzése után, regisztrációs folyamatot követően válhatnak felhasználókká.

A digitalizált anyagok máris új munkába állnak

A KDK-repozitóriumok fejlesztésének részeként a KDK 2020-ban elkezdett foglalkozni a mesterséges intelligencia kínálta eszközök társadalomtudományos archívumokban történő hasznosíthatóságával. A lehetőségek feltérképezéséből fejlesztési irány lett, amelynek megvalósításában az újonnan digitalizált interjúk is szerepet kapnak. A projekt célja a TK KDK két archívuma, a KDK és a 20. Század Hangja Archívum és Kutatóműhely állományaiban a nagyobb fokú áttekinthetőség és a komplexebb kereshetőség biztosítása a kutatási dokumentumok, elsősorban az interjúk metaadatainak gazdagításával. Az új metaadatok egységes tárgyszólistából származó tárgyszavak társításával, illetve a dokumentumokban szereplő névelemek (személynevek, földrajzi nevek, intézménynevek) és időelemek (dátumok, korszakok, ünnepek) azonosításával kerülnek a korábbiak mellé – mesterséges intelligenciát is használó, tanuláson alapuló algoritmusok segítségével.

A KDK a célok eléréséhez a Számítástechnikai és Automatizálási Kutatóintézettel (SZTAKI) együttműködve a Mesterséges Intelligencia Nemzeti Laboratórium (MILAB) kutatási programjának finanszírozásával, 2020-ban pilot projektet indított, majd 2021-ben, azt folytatva, komoly előkészítő és fejlesztő munkába fogott. A munka során a KDK és a SZTAKI munkatársai a manuális és a gépi szövegfeldolgozás

módszertanának kidolgozásával, egységes társadalomtudományos tárgyszólista meghatározásával, többféle kulcs- és tárgyszavazási módszer kipróbálásával, tesztelésével és validálásával, annotáló program segítségével történő manuális interjúkódolással, majd gépi tárgyszavazási módszerek alkalmazásával végezték a fejlesztőmunkát.

A feladathoz több különböző kutatási gyűjteményünk anyagát is bevontuk, részben a tanításhoz, részben a kapott eredmények ellenőrzéséhez. A folyamat során kiválasztott (legmegfelelőbbre értékelt) algoritmussal a tanulóhalmaz kézi kódolása alapján valamennyi digitálisan elérhető interjúnk anyagához tárgyszavakat rendeltetünk, illetve azokban névelemeket, időelemeket jelöltetünk ki, ezzel fejlesztve és egységesítve repozitóriumaink metaadatkészletét és bővítve a bennük való keresési lehetőségeket.

A módszer alkalmazása annál hatékonyabb, minél nagyobb és minél változatosabb szövegkorpuszon tudjuk tanítani és tesztelni a SZTAKI kutatói által fejlesztett algoritmust, így a munkába több típusú, különböző módszerrel készült, eltérő témájú és mélységű interjúkat vonunk be. A HRDA pilot projekt keretein belül digitalizált két kutatási anyag nagymértékben hozzájárul a MILAB-fejlesztés sikeréhez, speciális témáikkal, nyelvezetükkel, az azokban használt szakkifejezésekkel, a csak azokban előforduló név- és időelemekkel.

A KDK MILAB projektje 2022 őszén érkezett ötödik fázisába. Az eredmények, az új tárgyszó-hozzárendelések, a névelem- és időelem-kiemelések hamarosan a KDK repozitóriumainak keresőfelületein is megjelennek majd (a többi metaadatot kiegészítve), ezzel támogatva a kutatók és a nagyközönség kutatómunkáját, az archívumokban való kiigazodást, az adott kutatási kérdések szempontjából releváns szövegek, szövegrészek megtalálását.

Összegzés

A HRDA és az MTA KIK pályázatának köszönhetően két értékes kutatási anyaggal bővült a TK KDK archívuma, amelyek így elérhetőek és kutathatóak lettek. A projekt a KDK munkatársai számára lehetővé tette különböző új, korábban nem használt technikák kipróbálását, amelyek egy részét azóta is alkalmazzuk archívumaink napi gyakorlatában.

A munkának köszönhetően nehezen hozzáférhető analóg kutatási anyagok váltak néhány hónap alatt mesterséges intelligenciát alkalmazó kutatás szövegkorpuszának részévé.

Irodalomjegyzék

Gyurgyík 1999
Gyurgyík László, *A szlovákiai vegyes házasságok demográfiai vonatkozásai 1949-től napjainkig*, In.: Fórum Társadalomtudományi Szemle 1991/1 pp. 5–18.
https://epa.oszk.hu/00000/00033/00001/gyurgyik.htm

Kovách 2010
Kovách Imre, *A jelenkori magyar vidéki társadalom szerkezeti és hatalmi változásai*. Akadémiai nagydoktori thesis, MTA Politikai Tudományok Intézete
http://real-d.mtak.hu/296/

Kovách-Megyesi 2018
Kovách Imre – Megyesi Boldizsár, *A vidék harminc éve. A magyar vidék alakulása az erőforrások, a társadalmi tőke és fejlesztéspolitikai változásainak tükrében*, In: Erdélyi Társadalom 16 (1), 2018,
https://doi.org/10.17177/77171.209.

Kovács 2003
Kovács Éva, *A „házassági piac" alakulása Komáromban (1900–1940)*, In.: K. Horváth Zsolt – Lugosi András – Sohajda Ferenc (szerk.): Léptékváltó társadalomtörténet, Hermész Kör – Osiris: Budapest, pp. 366–394.
https://www.academia.edu/7769693/A_h%C3%A1zass%C3%A-1gi_piac_alakul%C3%A1sa_Kom%C3%A1romban_1900_1940_