# VALÓS TÉRBEN – AZ ONLINE TÉRÉRT

## Networkshop 31: országos konferencia

2022. április 20–22.
Debreceni Egyetem

Szerkesztette: Tick József, Kokas Károly, Holl András

HUNGARNET Egyesület
Budapest, 2022

NETWORKSHOP 2022

A kötet megjelenését támogatta az
Energiaügyi Minisztérium

Szerkesztette: Tick József, Kokas Károly, Holl András

Tipográfia és tördelés: Vas Viktória

Networkshop
2022. április 20–22. Debreceni Egyetem, konferencia előadásainak közleményei

ISBN 978-615-82243-0-7
DOI: 10.31915/NWS.2022

Kiadja a HUNGARNET Egyesület
az MTA Könyvtár és Információs Központ közreműködésével
Budapest
2022

Borítókép: freepik.com

## TARTALOMJEGYZÉK

DOI: 10.31915/NWS.2022.20

# Szociológia, kutatási adatok, mesterséges intelligencia: lehetőségek és tapasztalatok

Egyed-Gergely Júlia, Vajda Róza, Gárdos Judit, Horváth Anna, Meiszterics Enikő
*Társadalomtudományi Kutatóközpont (TK KDK)*
egyed-gergely.julia@tk.hu

Micsik András, Martin Dániel, Marx Attila, Pataki Balázs, Siket Melinda
*Számítástechnikai és Automatizálási Kutatóintézet (SZTAKI DSD)*
micsik@sztaki.hu

**Abstract**

Social science, research data and artificial intelligence: options and lessons learned

Our aim was to facilitate the research of collected interview texts at the Research Documentation Centre of the Centre for Social Sciences. Currently it is difficult to find the parts related to specific topics or obtain a thematic mapping of lengthy interviews and collections. To this end, we identified and tested the most promising NLP tools supporting the Hungarian language. Furthermore, a suitable and domain-oriented taxonomy had to be created for the classification of available texts. Our work and experiences are described in this paper, including the adaptation of the European Language Social Science Thesaurus, the building of a gold standard for classification, the comparison of various automated indexing methods, and finally the implementation of a researcher tool supporting custom visualizations and faceted search.

**Keywords:** natural language processing (NLP), named entity recognition, extreme classification, exploratory UI, text visualization

## Bevezetés

A Társadalomtudományi Kutatóközpont Kutatási Dokumentációs Központja (TK KDK) a TK négy intézetében (Jogtudományi Intézet, Kisebbségkutató Intézet, Politikatudományi Intézet, Szociológiai Intézet) folytatott kutatások során keletkezett dokumentumokat (interjúkat, kérdőíveket, kutatási terveket stb.) gyűjti össze, rendszerezi és teszi elérhetővé online felületén az anyagok másodfelhasználásában, illetve tudománytörténeti és/vagy módszertani vizsgálatában érdekelt kutatók számára. 2019 óta a KDK az 1960 és 2010 közötti hazai kvalitatív szociológiai kutatások szétszórtan heverő anyagainak (elsősorban interjúleiratoknak, emellett fotóknak, rajzoknak, videóinterjúknak, jegyzeteknek, tanulmánytöredékeknek stb.) gyűjtésével, digitalizálásával és kutatási célokra való feldolgozásával foglalkozó 20. Század Hangja Archívum és Kutatóműhelyt is magába foglalja. A két online dokumentumtár több tízezer digitális fájlja ingyenesen hozzáférhető, de az anyagok egy része regisztrációköteles.

Az állomány folyamatos bővülése és az átláthatóság kívánalma egyre nagyobb kihívások elé állítja az archívumokat üzemeltető KDK munkatársait. A dokumentumok alapadatait rögzítő metaadat-struktúra ugyanis csak korlátozottan teszi lehetővé az anyagok tematikus áttekintését, kereshetőségét, rendszerezését, ráadásul a két archívum metaadatrendszere különbözik egymástól. Mindkét archívum alkalmaz ugyan tárgyszavakat, azonban nem szöveg- illetve dokumentumszinten, illetve nem egységes szókészletből, és nem azonos módszerekkel történik a hozzárendelés. Az eddigi általános gyakorlat szerint kutatásszinten

(egy kutatáshoz rendszerint több dokumentum is tartozik) szerepel néhány – a 20. Század Hangja Archívumban 3-3, a KDK anyagai esetében sem sokkal több – esetlegesen meghatározott tárgyszó. Mindezek miatt az archívumokat látogató kutatóknak invenciózusan saját módszerekhez kell folyamodniuk, hogy a céljaiknak megfelelően feltérképezhessék a szövegállományt és kiválaszthassák a munkájukhoz szükséges dokumentumokat.

A tartalomban történő böngészést szövegszerű keresési funkció szolgálja, a két archívumban külön-külön. Ez persze nagyon hasznos, ha egészen specifikus fogalom vagy helyszín érdekli a kutatót. Amint azonban általánosabb jelenségre, tématerületre kíváncsi, kénytelen először "lefordítani" azt különféle széles körben használt kifejezésekre, amelyekről feltételezi, hogy szó szerint szerepelhetnek a szövegekben. Márpedig javarészt strukturálatlan interjúkból álló, meglehetősen heterogén szövegállomány esetében nem kis munka és fejtörés különféle fogalmakkal „körüllőni" egy-egy témát. Például, ha a társadalmi mobilitás kérdése foglalkoztatja az illetőt, akkor kezdheti olyan szavakkal, mint „szegénység", „pénz", „oktatás", és folytathatja még hosszan a sort. Azonban nem remélheti, hogy vaktában eltalálva az összes vonatkozó kifejezést rálel valamennyi, számára releváns dokumentumra. Ugyanígy, ha mondjuk a főváros története érdekli, nem elégedhet meg pusztán a „Budapest" szóra kereséssel, hanem végig kellene próbálnia valamennyi helynevet a Margit hídtól Ferencvárosig, ami kész lehetetlenség. De jogos kutatói igény lehet az is, hogy valaki egyszerre lássa át, melyek a leggyakrabban említett helyek, szereplők, intézmények, időszakok, vagy tudja meg, melyik kutató mikor, mivel foglalkozott.

Az anyagok „feltárása", az állomány áttekinthetővé tétele mint sürgető feladat és a feladat megoldásával kecsegtető technológiai újítások elérhetővé válása együtt adta a projektünk alapgondolatát. Kézi tárgyszavazás ilyen nagy szövegkorpuszon óriási mértékű humánerőforrást igényelne, ezért a lehetőségeink ismeretében ez fel sem merült. Mesterséges intelligenciára épülő algoritmusokat ezidáig ugyan más, egyszerűbb problémák megoldására használtak a társadalomtudományokban, mégis úgy gondoltuk, ez az eszköz alkalmas lehet a céljaink eléréséhez.

### A tárgyszókészlet és a tanítóhalmaz létrehozása

A két archívum kutatási anyagainak gépi tárgyszavazását többfázisos manuális előkészítés előzte meg. A mesterséges intelligenciát is használó algoritmusok betanításához, illetve a kapott eredmények validálásához tanítóanyagra volt szükség. Ehhez létre kellett hozni egy egységes tárgyszókészletet, meg kellett határozni a tanulókorpuszt, tesztelni rajta a tárgyszókészletet, majd előállítani a gold standardot. E lépések több körben, egymást javítva követték egymást.

### Tárgyszókészlet

Az archívumok kutatási anyagainak tárgyszavazásához strukturált tárgyszókészletre van szükség (Molnár, 2016). Magyar nyelvű társadalomtudományos tárgyszókészlet a munka kezdetén nem állt rendelkezésre – a speciális feladat miatt ráadásul olyan készletre volt szükség, amelyik elég részletes ahhoz, hogy lefedje a két archívum anyagainak témáit, viszont kellően szűk ahhoz, hogy gépi módszerekkel dolgozni lehessen vele, mindemellett később a két archívum közös keresőjének alapjául is szolgálhat. E feltételek egymással ellentétes irányba hatnak: az első és a harmadik minél részletesebb szókészletet, míg a második inkább átfogó szakszavakat igényel. Az egyensúly megtalálása az egyik oldalról kompromisszumokkal jár, míg a másik oldalról kemény korlátokba ütközik. Fel kellett tehát adni az aprólékos, részletekbe menő tárgyszó hozzárendelést, és alkalmazkodni a kódolás korlátos lehetőségeihez.

Az általános magyar tezauruszt (OSZK Köztaurusz), különböző szakkönyvek tárgyszavait és különböző nemzetközi társadalomtudományos szókészleteket átnézve kiindulásul a CESSDA[1] (Consortium of European Social Science Data Archives) ELSST[2] (European Language Social Science Thesaurus) angol nyelvű tezauruszára esett a választás. Emögött az a megfontolás állt, hogy a kutatási területen bevett nemzetközi szókészlet választása egyfelől elősegíti egy hiánypótló magyar társadalomtudományos tárgyszókészlet előállítását, másfelől lehetőséget biztosít nemzetközi projektekbe való bekapcsolódásra, valamint a két archívum nemzetközi láthatóságának növelésére. A munka során maga a fordítás önálló projektté vált, a TK és a CESSDA együttműködése eredményeként a magyar változat 2022 szeptemberében kerül az eddig 14 nyelven elérhető szókészlet mellé és lesz hozzáférhető az ELSST honlapján.

A saját tárgyszókészlet előállításához először a valamivel több mint 3300 kifejezésből álló tezaurusz magyarra fordítására került sor. Az eredeti angol nyelvből a SZTAKI kutatói által létrehozott SZTAKI-s és más online szótárak segítségével gépi, majd ezt követően manuális módszerrel ültettük át magyar nyelvbe a kifejezéseket. Ezután a fordítás többszempontú (jogi és nyelvészeti) lektorálása következett, majd az eredményt a projekthez igazítottuk (Albaugh et al, 2013, Albaugh et al, 2014). A fordítás olyan nem várt dilemmákat vetett fel, mint például az angol nyelv alapvetően többesszámot használó formájának átvétele vagy a kifejezések egyesszámúsítása, a magyar nyelvben nem használt (például speciális jogi) kifejezések megfelelő fordításának megtalálása, vagy a nem társadalomtudományos kifejezések problémája. A munka során újragondoltuk a nemzetközi szókészlet bizonyos elemeit, amivel támogattuk annak megújulását – a konzorcium folyamatosan karban tartja a tárgyszókészletet, az éves frissítés részben a magyar fordítás tapasztalatai alapján történik.

A szókészlet testreszabása során többkörös szűkítéssel, tömbösítéssel, pontosítással és kiegészítéssel elkészült egy 220 elemű, a gépi tárgyszavazáshoz alkalmasnak tűnő tárgyszólista. A lista kialakításánál szempont volt a relevancia (társadalomtudományi fogalmak kiválasztása a magyar szociológia témáinak, alapfogalmainak figyelembevételével), a lefedettség (minden fontos társadalomtudományi tématerület képviselete), az arányosság (az archívumok súlypontjaihoz való igazodás), a diszjunktivitás (jól elkülönülő fogalmak kiválasztása) és a mennyiség (gépi algoritmusok számára befogadható számú fogalom). Az eredeti lista drasztikus szűkítése mellett szükség volt az új kifejezések felvételére is, elsősorban a magyar szociológia és a magyar történelem témáinak megragadhatósága végett. A tárgyszólistát háromszintű hierarchikus rendszerbe rendeztük – támogatandó a manuális és gépi kódolást, valamint a KDK keresőjének fejlesztését.

### Tanítóhalmaz

A tanítóhalmaz létrehozásához a tanulókorpuszt a két archívum interjúkat tartalmazó kutatási anyagaiból kiválasztott 21 (735 oldalnyi) interjú adta. A kiválasztásnál szempont volt, hogy különböző témájú (a szociológiai szakma nagyjaival készült élettörténeti interjúk, börtönviseltekkel folytatott beszélgetések, magukat romáknak vallókkal készített interjúk, a Kádár-korszak emlékezetével kapcsolatos kutatás során készült anyagok, stb), különböző hosszúságú, eltérő módszerrel készült (narratív interjú, félig strukturált interjú, fókuszcsoportos beszélgetés), más-más gyűjteményekből származó és mindkét archívumot megjelenítő anyagok kerüljenek a korpuszba.

A kiválasztott interjúk manuális annotálását összetett feladatként határoztuk meg: a tanítóhalmaz kialakításához az annotátoroknak kulcs- és tárgyszavakat[3] kellett rendelniük

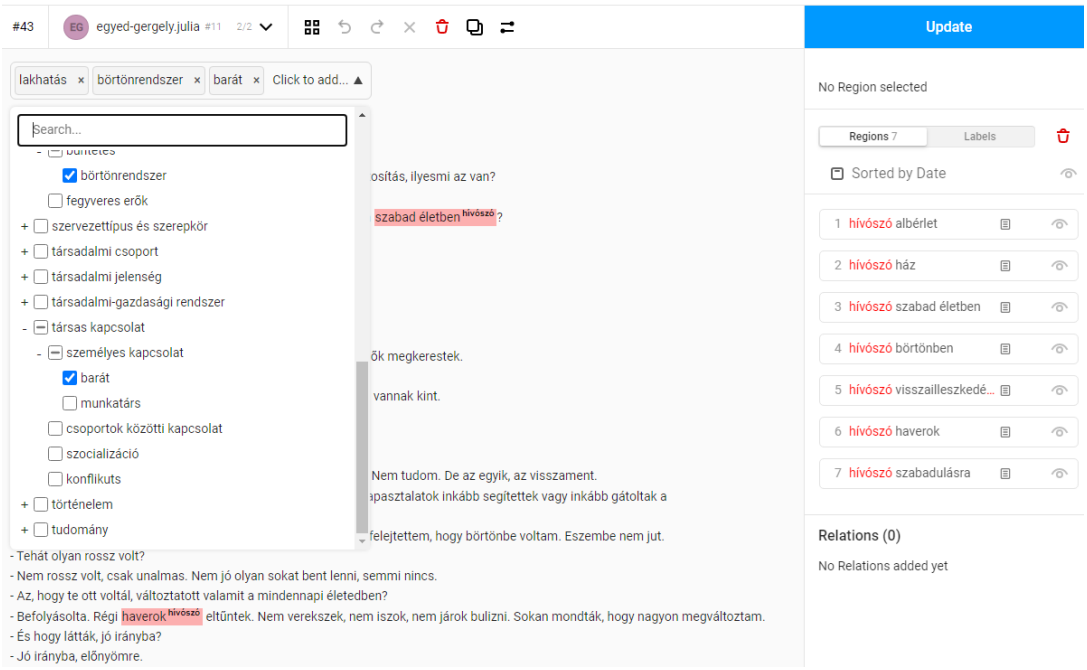
1 https://www.cessda.eu/

2 https://thesauri.cessda.eu/elsst/en/index

3 Míg a kulcsszavak a gépi annotálást hivatottak segíteni, addig a tárgyszavak feladata az interjú metaadatainak bővítése.

a szakaszokra bontott interjúszövegekhez. A hozzárendelés célja kettős volt: a tanítóanyag elkészülése mellett a saját tárgyszókészlet validálására, korrigálására is szolgált.

Az annotálást szakértők végezték, előzetesen kialakított irányelvek alapján (Molnár, 2016). A nagy mennyiségű tárgyszó és az alapvetően strukturálatlan szövegek miatt komoly kihívást jelentett az egységesség (Balázs-Sebők, 2016). Az annotálás Label Studio[4] felületen zajlott (1. ábra), amelyet a saját tárgyszókészlet betöltésével és az interjúfeldolgozáshoz alkalmas struktúra és megjelenítés kialakításával a SZTAKI kutatói a projekthez igazítottak. A szakaszokra bontott interjúkat először két-két annotátor egymástól függetlenül kódolta, majd az egyezés mértékétől függően vagy az eredeti kettő, vagy harmadik, független annotátor határozta meg a gold standardot.




*1. ábra. Annotálás közben a Label Studio felülete és a hierarchiába rendezett tárgyszókészlet*

Az egyezés meghatározását a projekt jellegéhez igazodva, saját elvek alapján alakítottuk ki. Egyezőnek tekintettük a tárgyszó-hozzárendelést, ha a tárgyszókészlet hierarchiájában a hozzáadott tárgyszóhoz tartozó, a legfelső szinten lévő kifejezés azonos volt. Azaz például, ha egy szakaszhoz az egyik annotátor a „szokások és hagyományok" tárgyszót, míg a másik a „kulturális esemény" tárgyszót adta, akkor ezt egyezésként fogadtuk el, ugyanis a hierarchiában mindkét kifejezés a „kultúra" tárgyszó alatt helyezkedik el. A két annotátor, bár más tárgyszót választott, ugyanazt a nagyobb területet jelölte velük.

Az egyezés kívánt mértékét szintén a feladat sajátosságai szerint határoztuk meg. 30% alatti egyezésnél harmadik annotátor, 30% feletti egyezésnél az eredeti két annotátor döntött a gold standardról. A szakirodalom ennél szigorúbb egyezést szokott elvárni, a speciális alkalmazás azonban speciális feltételek kialakítását igényelte és tette lehetővé. A „hagyományos" annotálási feladatokhoz képest több szempontból is különlegesnek ítéltük meg a helyzetet. Az interjúkat nem bonthattuk mondatokra vagy nagyon rövid (és ennek megfelelően várhatóan egyértelmű és könnyebben megfogható tartalmú) egységekre, ebben az esetben túlságosan sok, és gyakran irreleváns tárgyszót kaptunk volna. Egy-egy szövegegység általában több témát ölelt fel, az annotátoroknak ezek közül kellett megjelölnie a relevánsabbakat, a tárgy megjelölésén túl így a kiválasztás is feladat volt. Egy adott szakaszhoz nem egy (a legrelevánsabb) tárgyszót, hanem többet is lehetett rendelni, a hangsúlyosan

4 https://labelstud.io/

megjelenő témák számától függően. Ez nagyban nehezítette az egységes eredmény elérését, komoly kihívás elé állítva a munkafolyamat kialakítóit és az annotátorokat (Albaugh et al, 2014). A munka jellegéből adódóan a tárgyszavak egymáshoz képest nem kizáróak, nem is lehetnek azok (mint például a pozitív/negatív/semleges kódolásánál), így két kódoló által adott különböző tárgyszavak egyformán relevánsak lehetnek akár még akkor is, ha még csak nem is ugyanahhoz a nagyobb tématerülethez tartoznak. Az egy szakaszban megjelenő különböző témák közül a relevánsak kiválasztásának módja nehezen egységesíthető. Egy szakaszban megjelenhet egymás mellett például a „térbeli elhelyezkedés", a „jövőkép" és a „fogyasztás" mint fontosabb téma, ezek közül bármelyik és bármelyek kombinációjának kiválasztása releváns döntés lehet a kódoló részéről. Végül a korpusz jellege (gyakran leginkább a köznyelvhez hasonlító interjúszövegek) is indokolta a speciális – a szokásosnál megengedőbb – irányelvek kialakítását.

A gold standard és ezzel a tanítóhalmaz előállítása után a KDK és a 20. Század Hangja Archívum nagy mennyiségű interjúanyagának gépi annotálása következett, 368 interjú bevonásával, majd az eredmények manuális validálása véletlenszerűen kiválasztott interjúk ellenőrzésével.

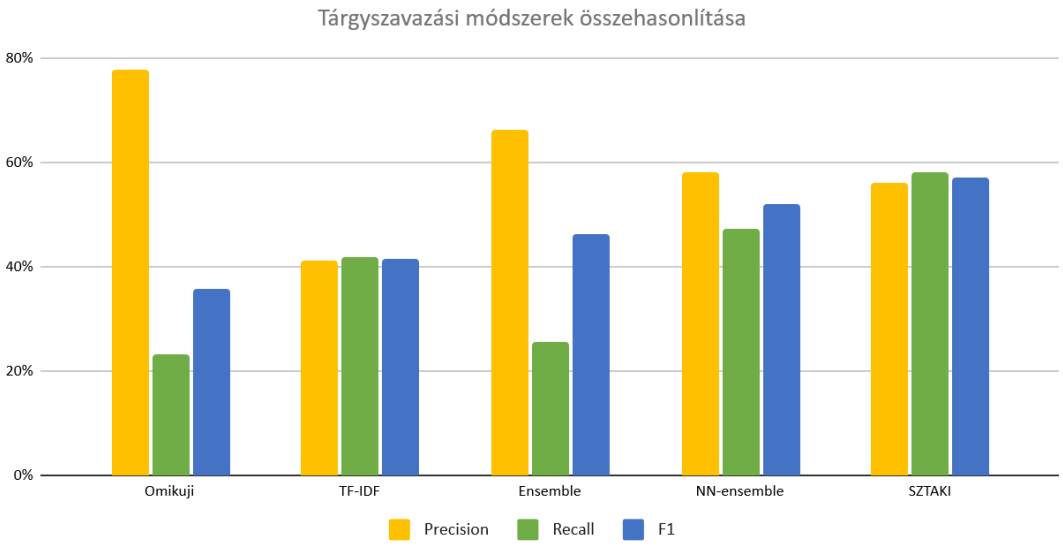
### Automatikus tárgyszavazás

Az automatikus tárgyszavazási feladat főbb informatikai jellemzői: 2-5000 karakternyi szövegrészekhez (szakaszokhoz) kell néhány jellemző tárgyszót kiválasztani egy háromszintű tárgyszó taxonómiából. Ez egy elég speciális feladat, amelyhez a hagyományos klasszifikációs MI algoritmusok nem használhatók a nagy taxonómia miatt, a topic modelling megoldások pedig azért, mert előre adott a taxonómia. Szerencsére kifejezetten erre a célra jött létre az annif szoftvereszköz (Suominen et al. 2022), amellyel tízféle módszer közül lehet választani ilyen jellegű tárgyszavazáshoz. A módszerek között van tanuló és statisztikus, illetve kombinálni is lehet a módszereket. Az annif wiki[5] oldalán megtalálható az egyes módszerek rövid leírása és a megalapozó tudományos cikkek hivatkozásai. Az összes módszert kipróbáltuk a gold standard korpuszunkkal, és a legjobban teljesítők eredményét tüntettük fel a 2. ábrán. A vizsgált módszerek közül a legjobb eredményt elért NN-ensemble az Omikuji és a TF-IDF tárgyszavait 3:1 arányban kombinálja.

Kicsivel ennél is jobb eredményt ért el saját fejlesztésű módszerünk, amely a tanítóhalmaz készítése során a tárgyszavakhoz gyűjtött kulcsszavak használatán alapul. Itt a textacy[6] és a legfrissebb huspacy (Orosz et al. 2022) segítségével kulcsszavakat gyűjtöttünk a szövegrészhez, és a kulcsszavakhoz tartozó tárgyszavak statisztikája alapján alakítottuk ki a végső tárgyszólistát. A textacy előnye, hogy többféle kulcsszavazási módszer is választható. Ezek közül számunkra a yake adta a legjobb javaslatokat, ezt szubjektív vizsgálati módszerekkel sem volt nehéz eldönteni.

5 https://github.com/NatLibFi/Annif/wiki

6 https://github.com/chartbeat-labs/textacy



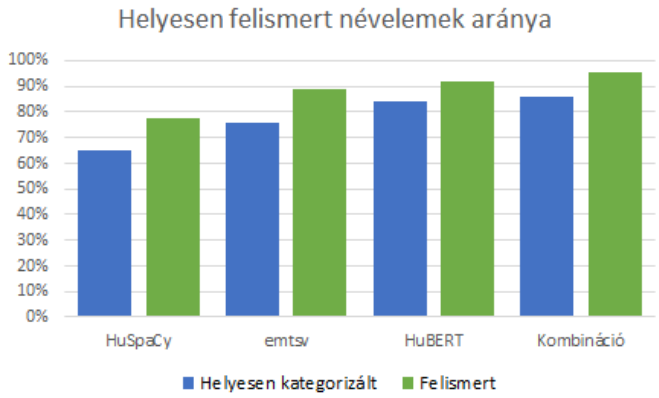

*2. ábra. A különböző tárgyszavazási módszerek eredményeinek összehasonlítása*

Az első és második helyezett módszer fenntarthatóságát érdemes megvizsgálni. Például, mi történik, ha meg kell változtatni a tárgyszórendszert? Az első helyezett módszer esetében a megváltozott tárgyszavaknál át kell alakítani a társított kulcsszólistát, és validálni kell egy kialakított kisszámú mintán. Ezután újra kell futtatni a tárgyszavazást az interjúkon (vagy egy részükön), amely időigényes lehet, mivel a szöveget nyelvileg elemezni kell (szótövezés, szófaj meghatározás stb.).

A második helyezett módszer esetében kissé bonyolultabb a helyzet: a módszert újra be kell tanítani, ehhez pedig ki kell bővíteni a tanulóhalmazt a változott tárgyszavak elegendő számú előfordulásával. Ez egy iteratív folyamat, amelynek során, ha az eredmények már elég jók, akkor lehet újrafuttatni a tárgyszavazást az interjúkon.

### Névelemek felismerése, összekapcsolása

Másik fontos célunk volt kinyerni az interjú szövegekből az említett földrajzi, személy- és egyéb tulajdonneveket. A névelem-felismerés vizsgálatához kialakítottunk egy több mint 3500 névelemet tartalmazó minikorpuszt az interjúkból, és azon futtattunk három névelem-felismerőt: a már említett huspacy-t, a korábban készült emtsv emBert modulját (Nemeskey 2020), valamint a huBERT-et (Nemeskey 2021). Utóbbit az időközben elkészült NYTK-NerKor korpuszon (Simon 2021) tovább tanítottuk, és mivel ez a legbővebb és legfrissebb névelem-felismerési korpusz, talán ennek tudható be, hogy ezzel értük el a legjobb eredményt. Még ez sem volt azonban elég jó, a névelemek majdnem tizede kimaradt. Amikor a három módszer különbözőképp ítélt meg egy névelemet, egyszerű szabályok alapján (pl. találtunk-e hozzá Geonames azonosítót) kombinált eredményt alakítottunk ki, amelyhez már a wikifikálás kimenetét is felhasználtuk. Ezen a területen is van még fejlődési lehetőség, mivel a névelemek 5%-a rejtve maradt, 9%-a pedig rossz kategóriába került.

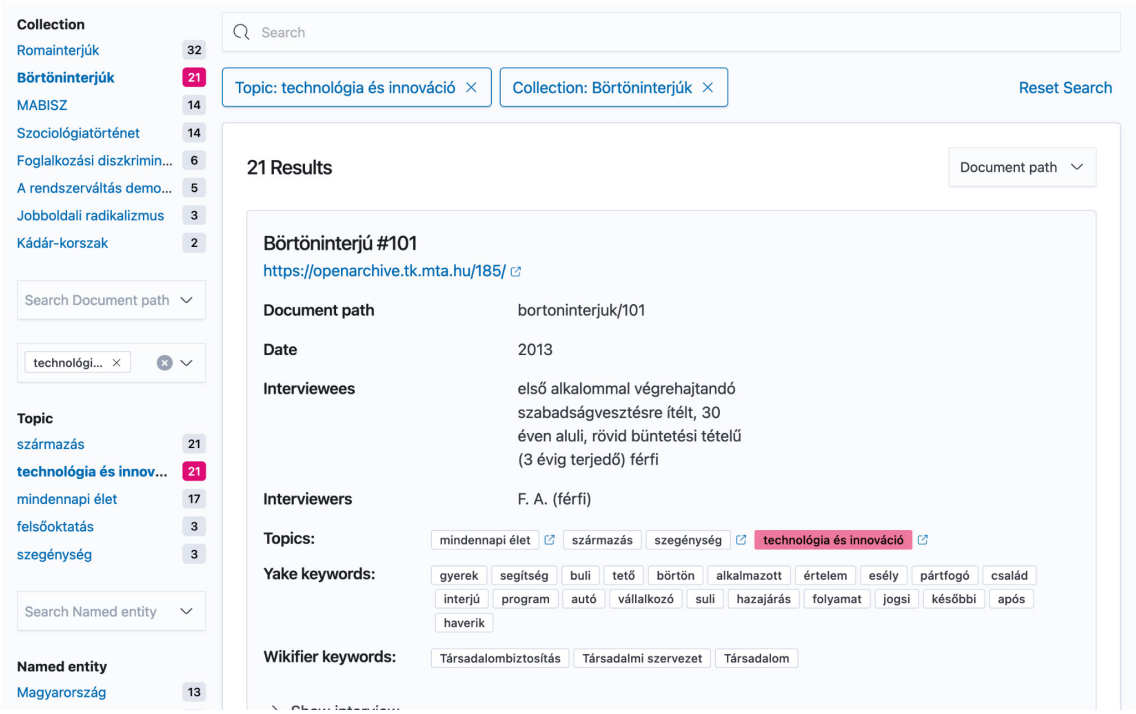



*3. ábra. A különböző nyelvi elemzők névelem felismerési eredményei*

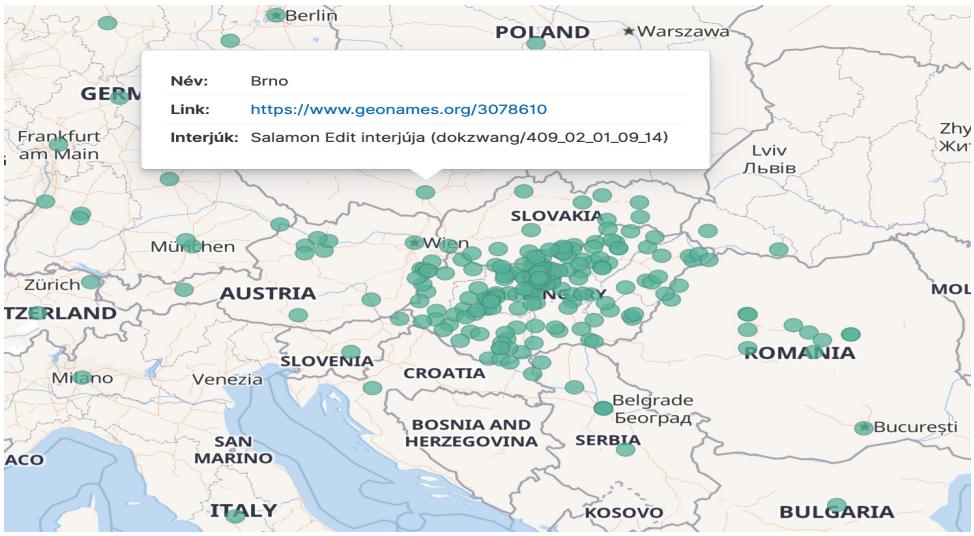
A felismert névelemekre rákerestünk a Wikidata-ban azért, hogy a következő fejezetben bemutatott interaktív interjú olvasóban linkeket tudjunk felajánlani az egyes nevek említéseihez. A Wikidata aggregálja sok más regiszter azonosítóját, így a hivatkozások nagy részét a Wikidata-ból meg lehet szerezni (pl. PIM, VIAF, Geonames). Viszont az egyértelműsítés (*disambiguation*) terén sok munka lenne még: a szövegekben előforduló közgazdászok, írók, tudósok helyett rendszeresen azonos nevű focisták és más celebritások linkjeit kaptuk első helyen a Wikidata-tól. Pár, a tudásgráfból lekérhető alapvető jellemzőt megpróbáltunk figyelembe venni a választáskor, mint például az entitás típusa vagy a személy születési éve, de ez sem volt elegendő az eredmények jelentős javulásához.

## Kutatói munkafelület

Az összes interjú nyelvi elemzésével előállt egy adatbázis, amelyet a kutatók számára felfedezhetővé akartunk tenni. Ezért nyílt forráskódú komponensekből összeállítottunk egy fazettás keresőt, amelyben a teljes szöveges keresés mellett változatos szűrési lehetőségek is vannak (4. ábra). Az ábrán középen a kék keretes dobozokban látható a két aktív szűrőfeltétel, illetve a szűrt elemek piros háttérrel is meg vannak jelölve.




*4. ábra. Fazettás keresés az interjú-gyűjteményekben*




*5. ábra. Az interjúkban említett földrajzi nevek térképes megjelenítése*

A találati listában az interjúk főbb metaadatai, valamint az automatikus elemzés interjúra összesített leggyakoribb tárgyszavai, kulcsszavai és Wikidata oldalai szerepelnek. Az interjú teljes szövegét is meg lehet itt nyitni olvasásra, ahol a szövegben azonosított névelemek mellett kis információs panel jelenik meg a megfelelő Wikidata, Geonames, VIAF stb. oldalakra mutató linkekkel. Az egyedi vizualizációk készítéséhez egy Kibana[7] felületet is integráltunk a kereső mellé, amellyel számtalan grafikontípusban jeleníthetjük meg a kiválasztott adattartalmat. Az 5. ábrán példaként egy térképes vizualizációt mutatunk be, amely az interjúkban említett földrajzi helyeket prezentálja, illetve megmutatja azt is, hogy mely interjúkban történik az említés.

## Összefoglalás

A rendelkezésre álló nyílt elérésű szoftverekkel viszonylag könnyen lehet gazdag, felfedező, áttekintő jellegű (exploratív) kutatói felületeket létrehozni. Ez azért is szükséges, mert nagy mennyiségű értékes kutatási anyag létezik és keletkezik szöveges vagy hangfelvétel formájában, amit valamilyen előzetes feldolgozás nélkül nem lehet áttekinteni. A projekt során kipróbáltuk a legnépszerűbb és legfrissebb magyar és nyelvfüggetlen nyelvi feldolgozó eszközöket abból a szempontból, hogy mennyire használhatóak a konkrét kutatási célra a feldolgozás eredményei. Ehhez nélkülözhetetlen valamilyen viszonyítási alap, ám ennek előállítása meglehetősen erőforrásigényes. Meg kell jegyezni, hogy a szótövezés és névelem-felismerés magyar nyelven még mindig olyan mennyiségben hibázik, hogy az kutatói környezetben zavaró, és utólagos kézi javítást igényelhet. A másik kulcskérdés az, hogy milyen tárgyszórendszer a legalkalmasabb a tömeges, automatikus kategorizálásra. Ezen a téren egy európai szociológiai taxonómia hazai adaptációja mellett döntöttünk, bár valójában nem volt igazi alternatíva. Előremutató lenne a magyar nyelvű taxonómiákat is összegyűjteni és megfelelően formalizálni (pl. SKOS), hogy segítsünk a jövőben megbirkózni a hasonló kihívásokkal.

**Finanszírozás:** A publikációban bemutatott projektet, amelyet a TK és a SZTAKI valósított meg, az Innovációs és Technológiai Minisztérium és a Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében.

7 https://www.elastic.co/kibana/

**Bibliográfia**

ELSST – European Language Social Science Thesaurus. https://elsst.cessda.eu/

Albaugh, Quinn - Sevenans, Julie - Soroka, Stuart - Loewen, Peter John (2013): *The Automated Coding of Policy Agendas: A Dictionary-Based Approach.* Paper presented at the 6th annual Comparative Agendas Project (CAP) conference, Antwerp, June 27–29.

Albaugh, Quinn - Soroka, Stuart - Joly, Jeroen - Loewen, Peter John - Sevenans, Julie - Walgrave, Stefaan (2014): *Comparing and Combining Machine Learning and Dictionary-Based Approaches to Topic Coding*. Paper presented at the 7th annual Comparative Agendas Project (CAP) conference, Konstanz, June 12–14.

Balázs Ágnes – Sebők Miklós (2016): Névelem-felismerés, In.: Sebők M. (szerk): Kvantitatív szövegelemzés és szövegbányászat a politikatudományban, L'Harmattan, 2016, pp.51–61.

Molnár Csaba (2016): Szövegkódolás a gyakorlatban: kézi, géppel támogatott és gépi megoldások, In.: Sebők M. (szerk): Kvantitatív szövegelemzés és szövegbányászat a politikatudományban, L'Harmattan, 2016, pp. 24-36.

Suominen, O., Inkinen, J., & Lehtinen, M. (2022). Annif and Finto AI: Developing and Implementing Automated Subject Indexing. *JLIS.It*, *13*(1), 265–282. https://doi.org/10.4403/jlis.it-12740

Orosz György, Szántó Zsolt, Berkecz Péter, Szabó, Gergő, Farkas Richárd (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In XVIII. Magyar Számítógépes Nyelvészeti Konferencia.

Nemeskey Dávid Márk: Egy emBERT próbáló feladat. XVI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2020), Szeged, pp. 409–418.

Nemeskey Dávid Márk (2021): Introducing huBERT. XVII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2021). Szeged, pp. 3–14

Simon, Eszter; Vadász, Noémi. (2021) Introducing NYTK-NerKor, A Gold Standard Hungarian Named Entity Annotated Corpus. In: Ekštein K., Pártl F., Konopík M. (eds) Text, Speech, and Dialogue. TSD 2021. Lecture Notes in Computer Science, vol 12848. Springer, Cham. https://doi.org/10.1007/978-3-030-83527-9_19