https://doi.org/10.3311/tmt.13239

Repozitóriumi gyűjtemény mint adatkorpusz[1]

Holl András, Prószéky Gábor, Váradi Tamás, Laki László

Cikkünkben bemutatjuk a REAL repozitórium modern magyar nyelvű tartalmainak szövegkorpuszként való felhasználását, melyet az Akadémiai Könyvtár (MTA KIK) és a Nyelvtudományi Kutatóközpont (NYTK) közösen valósít meg az MTA „Tudomány a Magyar Nyelvért Nemzeti Programjának" keretében. „A magyar nyelv digitális támogatása a magyar tudományosság szolgálatában" alprogram 2026-ban fejeződik be, és az NyTK által fejlesztett neurálisháló-alapú nyelvtechnológiai eszköz tanításán és alkalmazásán alapul.

A REAL az egyik legjelentősebb hazai tudományos repozitórium, nyolc gyűjteménybe szervezett, 210 ezernél több tétellel, melyek túlnyomó többsége szabadon letölthető. A repozitóriumból havonta közel félmillió letöltés történik. Az új projekt keretében megkísérreljük a szöveges tartalmak egy részét szövegkorpuszként hasznosítani. Mindehhez az eddiginél alaposabban meg kell ismerjük saját gyűjteményünket, fel kell mérjük adattartalmát, és meg kell vizsgáljuk a dokumentumok és leíró adataik minőségét. Ezt követően törekednünk kell a leíró adatok javítására és kiegészítésére, még a nyelvtechnológiai eszközök alkalmazását megelőzően. A projekt eredményeképpen reményeink szerint a repozitóriumi adatok tovább javíthatóak majd.

A projekt része a szövegkorpusz bányászata: a szövegekben azonosítjuk az alapvető bibliográfiai információkat és segítségükkel mind az eredeti dokumentum leírását gazdagítjuk, mind más adatbázis (az MTMT) kiegészítéséhez felhasználjuk ezeket. Foglalkozunk a szövegek tématerületi osztályozásával is. Mindezen feladatok elvégzésében támaszkodhatunk a REAL-ban alkalmazott, nyílt forráskódú, szabadon használható EPrints szoftverre.

repozitórium, gyűjtemény, adatkorpusz, REAL, könyvtár

Bevezetés

A Magyar Tudományos Akadémia „Tudomány a Magyar Nyelvért Nemzeti Programja" egyik alprogramjának, „A magyar nyelv digitális támogatása a magyar tudományosság szolgálatában" projektnek a keretében a Nyelvtudományi Kutatóközpont és az MTA Könyvtár és Információs Központ a REAL repozitóriumrendszerben tárolt, modern magyar nyelvű tartalmak feldolgozásához látott hozzá. Az együttműködés keretében a felek egyrészt a szövegeket használják nyelvtechnológiai eszközök tanítására, másrészt a nyelvtechnológiai eszközök segítségével javítják és gazdagítják a Könyvtár adatbázisait, az MTMT-t és magát a REAL-t.

1 A Networkshop 2023 konferencián elhangzott előadás alapján

A projekt célkitűzései

Könyvtári oldalról a REAL statisztikáinak, metaadatainak és adatainak gazdagítása és javítása a minimális cél, és ebben már az előkészítés során sikerült előre lépni. A fő cél mindazonáltal a szövegbányászat. Névelem-felismerést (named entity recognition) kívánunk végezni a repozitóriumban tárolt modern magyar szövegekben, melynek eredményeképpen bibliometriai információkat remélünk kinyerni, ezáltal hivatkozásokat, a tudományfinanszírozókat érdeklő projekt- és nagyberendezés-azonosítókat, továbbá névterekbe köthető tulajdonneveket gyűjteni, amelyekkel az MTMT-t gazdagíthatjuk. A teljes szövegekből kinyerhető bibliográfiai adatokat a repozitóriumban tárolt metaadatok ellenőrzéséhez, esetenként kiegészítéséhez használhatjuk majd.

Holl, A., Prószéky, G., Váradi, T., Laki, L. *Repozitóriumi gyűjtemény mint adatkorpusz*, Tudományos és Műszaki Tájékoztatás, 70(2), p. 164–167, 2023. https://doi.org/10.3311/tmt.13239

Nyelvészeti oldalról a repozitóriumban található szövegek speciális tanítóanyagot szolgáltatnak nyelvmodellépítéshez. Mindezeken túl izgalmas kérdés, mennyiben járulhat hozzá a projekt a magyar tudományos szaknyelv ápolásához.

A REAL

A REAL jelenleg nyolc gyűjteményből álló, nyílt hozzáférésű repozitórium: a tárolt tartalmakat hosszú távon megőrző, és szabadon elérhetővé tevő digitális könyvtári rendszer. A gyűjtemények leírását a Könyvtár honlapja tartalmazza.[2] Az archivált dokumentumok száma meghaladja a 225 ezret. A "Ranking Web of Repositories" szerint az intézményi repozitóriumok között a Google Scholar adatai alapján a REAL alapgyűjtemény a 29. helyet foglalta el 2023 februárjában.[3] A repozitóriumi letöltések száma a pandémia óta havi fél millió körül alakul. A repozitóriumi digitalizálásról és a gyűjtemények leírásáról lásd még Holl et al. (2019) cikkét.

A projektben a gyűjtemények közül a törzsgyűjtemény (REAL), a könyveket tartalmazó REAL-EOD, a folyóiratgyűjtemény (REAL-J) és a disszertációgyűjtemények (REAL-D és REAL-PhD) modern, magyar nyelvű dokumentumaival foglalkozunk.

A REAL feltárása, a korpuszkészítés előzetes feladatai

Mind a REAL, mind a többi, hazai intézményi repozitórium, valamint a szöveges tartalmakat kínáló digitális könyvtárak alapvetően humán olvasóközönségnek készültek. Szövegminőség tekintetében az intézményi digitalizálás keretében a gyűjteménybe került tartalmak a leginkább egyenletes minőségűek, a szerzői önfeltöltés és a digitálisan született, szerkesztőségektől származó anyagok igencsak vegyesek lehetnek. A repozitóriumi feltöltések ellenőrzése jelent valamelyes minőségi kontrollt – de a cél itt is az olvasók kiszolgálása, nem a szövegbányászat. A projekt előkészítő fázisának első tanulsága a PDF állományokból kinyerhető szöveges réteg minőségével való szembesülés volt.

Eddig azt tudtuk megmondani, hány tétel van a repozitóriumban. Az előkészítés során pozitív meglepetés volt, milyen könnyű az EPrints szoftverből metaadatokat és szöveges dokumentumokat kinyerni gépi aratással. A munka eredményeképp más pontosabb információink vannak a szöveges tartalomról: hány dokumentum, hány szövegoldal, karakter, szó van az állományban lévő dokumentumokban, illetve mekkora a szöveges réteget tartalmazó dokumentumok aránya. Tudjuk, hogy a repozitóriumban találhatóak duplikált szövegek – például a cikkenként és folyóiratszámonként is archivált anyagok – de jelenleg nem rendelkezünk erről nyilvántartásokkal, sőt, statisztikákkal sem. Egy-egy dokumentumon belül is lehetnek olyan szövegrészek – például idézett szövegek, amelyek más dokumentumok részeként is archiválásra kerültek. A szöveg- és dokumentumismétlődések – azaz a duplumok – számbavétele is szerepel a projekt célkitűzései között.

A projekt magyar nyelvű szövegekkel dolgozik – de meg tudjuk-e mondani, milyen nyelvű egy dokumentum? Az EPrints dokumentumleíró sémájának része a 'nyelv' mező, azonban ez nem volt kivezetve sem a feltöltő űrlapokra, sem a dokumentumokat bemutató nézetekbe. Mint kiderült, a feltöltés során az EPrints a böngésző aktuális beállításai alapján töltötte ki a mezőt, aminek a dokumentum tényleges nyelvéhez ritkán volt köze. Az előkészítő fázis feladatai közé bekerült a nyelvet leíró metaadatmező megfelelő kitöltése. Ez százezres tételszámnál jelentős feladat! Azt az eljárást alkalmaztuk, hogy rendre a magyar, angol, majd további gyakori nyelvek feltételezésével a dokumentum címét teszteltük. Az adott nyelvi beállítással a címben található szavakat a hunspell helyesírás-ellenőrzővel vizsgáltuk meg, és ha a szavak több, mint a felét felismertük, a nyelvre vonatkozó feltevést elfogadtuk. Becsléseink szerint igen magas pontossággal sikerül így a nyelvi beállításokat rendezni. A projekt során további finomítást remélünk, annál is inkább, mert a tudományos publikációk gyakorta többnyelvűek.

A REAL-ban az EPrints szoftver által biztosított Library of Congress Classification (LCC) szakterületi beállításokat használjuk. Feltételezzük, hogy egyes feladatoknál a tanítóanyagokat és a feldolgozandó dokumentumokat érdemes lesz szakterületenként különválasztani. Ezért figyelmet kell fordítsunk arra, hogy az esetleg hiányszó szakterületi beállításokat pótoljuk, a meglévőket szükség esetén finomítsuk.

2 https://konyvtar.mta.hu/index.php?name=v_5_5
3 https://repositories.webometrics.info/en/institutional

Nyelvtechnológiai eszköztár

A szövegminőség javítása a PDF-ből kinyert digitális dokumentumokban nagymértékben elősegíti az információhoz való hozzáférést és a szövegfeldolgozás hatékonyságát. A minőség biztosítása érdekében a nyelvtechnológiai eszközök, mint például a természetesnyelv-feldolgozás (NLP), alapvető szerepet játszanak a szöveg strukturális és szemantikai értelmezésében. A szöveges adatok tisztításától kezdve, a szófajok azonosításán át, egészen a komplex szövegértési feladatokig, mint a szövegösszefoglalás vagy a kérdés-válasz rendszerek, az NLP-eszközök nélkülözhetetlenek a digitális szövegek minőségének biztosításához.

A neurális hálózatokra épülő technológiák, amilyen a mélytanulás is, forradalmasították az utóbbi években a nyelvtechnológiai alkalmazásokat. Az ilyen hálózatok képesek összetett mintákat felfedezni nagy adathalmazokban, beleértve a nyelvi adatokat is. A neurális hálózatoknak köszönhetően a természetesnyelv-feldolgozás képes volt elmozdulni a hagyományos, szabály alapú rendszerektől az összetettebb, adat alapú modellek felé, amelyek jobban képesek megbirkózni a nyelv használatának variabilitásával és összetettségével. Az olyan neurális hálózatok, mint a transzformer architektúrák (pl. GPT-3, BERT), képesek a szemantikai és a szintaktikai információk bonyolult kapcsolatainak megragadására, így javítva a PDF-ből kinyert szövegek minőségét és értelmezhetőségét (Yang et al., 2023).

A mesterséges intelligencia eszközei mellett más, egyszerűbb megoldásokat is szándékozunk alkalmazni. Lehetőleg nyílt forráskódú, szabad szoftvereket, esetlegesen az előkészítő fázisban vagy a repozitóriumi munkamenetekben alkalmazottaknál kifinomultabbakat választunk. Ilyenek lehetnek például az OCR-szoftverek, a nyelvfelismerő szoftverek, vagy a szövegtisztítás során alkalmazható és EPrints-hez illesztett aratószkriptek. Reményeink szerint egy közreadható, újrafelhasználható feldolgozási munkamenet és az ehhez szükséges eszköztár jön majd így létre.

Archiválási kihívások

Mi a repozitórium célközönsége? Ezidáig a humán olvasók voltak. Ebből az következett, hogy a gyarapítás – különösen az ellenőrzött szerzői önfeltöltés – során arra ügyeltünk, hogy emberi szemmel olvasható legyen a dokumentum. A szisztematikus és egyedi könyvtári digitalizálás során OCR-rel biztosított szöveges réteget kaptak a dokumentumok – mindennek a célja azonban csupán a teljes szöveges kereshetőség volt. A PDF-állományokból kinyert szövegek vizsgálata azt mutatta, hogy ezek a szövegek gyakorta rossz minőségűek: karaktertévesztések, rossz szegmentálás, sok karakterszemét nehezíti a majdani gépi feldolgozást. A digitálisan született szövegek sem problémamentesek, mert a tördelés, a szóelválasztások, vagy a nem törzsszöveghez tartozó elemek (lapszám, élőfej) nehezítik a feldolgozást. A szövegközi ábrák és táblázatok is problémát jelentenek mind az ún. born digital, mind a digitalizált szövegeknél. A problémák egy részével a modern, mesterségesintelligencia-alapú nyelvtechnológiai eszközök bizonyára megbirkóznak majd, mindazonáltal a szövegminőség monitorozására, esetlegesen a rosszabb minőségű szövegek kizárására is szükség lehet. Meg kell kíséreljük az esetleg repozitóriumba került nem kétrétegű PDF-állományok optikai karakterfelismertetését.

A projekt a repozitóriumokban és a kiadóknál alkalmazott eljárások javításában megnyilvánuló járulékos nyereségeket is eredményezhet. Az MTA KIK és a REAL esetében vizsgáljuk a saját kiadványainknál a PDF mellett a kiadványszerkesztőből kinyerhető tiszta szövegek elhelyezésének lehetőségét. Mint Mons (2005) találóan megjegyezte: "Why bury it first and then mine it again?"

A REAL gyarapítási politikája változik a szövegbányászati felhasználás megjelenésével. A szerzői jogi törvény 2021-es módosítása biztosítja a kutatóhelyek és kulturális örökségvédelmi intézmények által tudományos kutatás céljából végzett szöveg- és adatbányászat lehetőségét. Eddig nem gyűjtöttünk olyan anyagot, ami záros időn belül nem tehető szabad hozzáférésűvé. Mostantól indokolt esetben zártan is archiválunk folyóiratokat, kifejezetten a szövegbányászati felhasználás céljára. A szövegbányászati felhasználás megvalósulása szükségessé tette a publikus szolgáltatási megállapodás ilyen célú kiterjesztését is. A tágabb értelemben vett szövegbányászat lehetőségeiről lásd Holl (2015) cikkét.

A projekt kiterjesztésének lehetőségei

A feldolgozási munkamenet és eszköztár lehetővé teszi majd, hogy a korpuszépítést kiterjeszthessük további együttműködő intézményi repozitóriumokra is, első körben az EPrints szoftvert használókra. Mivel a REAL és a projektbe bevonandó további repozitóriumok gyarapodása folytatódik, a kialakult eljárásokat célszerű lesz időközönként újra megismételni. Remélhetőleg a projekt lehetőségei és elért sikerei segítenek majd a REAL gyűjteményeknek az eddiginél is szélesebb körből történő gyarapításában.

Irodalomjegyzék

Mons, B. *Which gene did you mean?*, BMC Bioinformatics 6, 142, 2005. https://doi.org/10.1186/1471-2105-6-142

Holl, A. *Szövegbányászat, adatbányászat, ismeretfeltárás*, Magyar Tudomány, 6, p. 680–685, 2015. Elérhető: http://real.mtak.hu/24408/

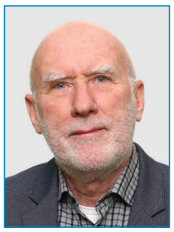
Holl, András; Horváth, H., Bilicsi, E., Nagy, E., Tömöry, P. *Folyóirat- és könyvdigitalizálás az MTA Könyvtárában. Lezárult az első szakasz*, Könyvtári Figyelő, 65(3), p. 375-382, 2019. Elérhető: http://real.mtak.hu/103280/

Yang Zijian, Gy., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, Á.,; Laki, L. J., Ligeti-Nagy, N., Vadász, N., Váradi, T., *"Jönnek a nagyok! BERT-Large, GPT-2 és GPT-3 nyelvmodellek magyar nyelvre"*, In: XIX. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2023. Elérhető: http://acta.bibl.u-szeged.hu/78417/

Beérkezett: 2023. május 23.

Holl András
informatikai főigazgató-helyettes
MTA Könyvtár és Információs Központ
E-mail: holl.andras@konyvtar.mta.hu

Prószéky Gábor
főigazgató
Nyelvtudományi Kutatóközpont
Email: proszeky.gabor@nytk.hu



Váradi Tamás
főigazgató-helyettes
Nyelvtudományi Kutatóközpont
E-mail: varadi.tamas@nytud.hu

Laki László János
tudományos munkatárs
Nyelvtudományi Kutatóközpont
Nyelvtechnológiai kutatócsoport
E-mail: laki.laszlo@nytud.hu