corpus. The Hungarian National Toponym Registry Program, established with the support of the Hungarian Academy of Sciences, has set itself the goal of collecting the Hungarian toponymicon of the Carpathian Basin in a comprehensive manner, presenting the names to both the professional community and the wider public after adequate documentation and the analysis of the individual names. The work exploring toponyms is carried out as part of a broad academic collaboration, primarily relying on universities of the region as its basis, where the staff of the program is trained and prepared specifically for this task. Online learning materials also facilitate the coherent implementation of the work. The program covers all Hungarian-speaking settlements of the region, but also toponym use in minority languages spoken in Hungary. In addition to the contemporary toponym corpus, it also takes into account toponym records from historical sources going back to the beginning of Hungarian written culture. Given the diversity of the scholarly uses of place names, the program is necessarily multi- and interdisciplinary in nature, ensured by the professional composition of the staff. The program planned to span a decade will publish toponyms in the form of online databases and printed books, thus updating and significantly expanding the overall corpus of the disciplines that use place names as source material.

Keywords: toponym research, toponym registries, toponym collection, Hungarian Academy of Sciences.

HOFFMANN ISTVÁN
Debreceni Egyetem

A magyar nyelv digitális támogatása a magyar tudományosság szolgálatában*

1. A kutatás előzményei. Ismeretes, hogy a globalizáció és a mindent átszövő digitális érintkezés korában a nyelvek nagy része, így anyanyelvünk is új kihívások elé néz. A fenyegetettséget nem nyelvünk kihalása jelenti, hanem az, hogy technológiai támogatás hiányában a nyelvek kiszorulhatnak a digitális térből (KORNAI 2013). Ha bizonyos digitális szolgáltatások magyarul például nem elérhetők, akkor a magyar emberek kénytelenek idegen nyelvet használni az anyanyelvük helyett. A magyar nyelvhasználat bizonyos mérvű visszaszorulását máris tapasztalhatjuk egyes nyelvhasználati területeken (JELENCSIK-MÁTYUS et al. 2022). Ezek közé tartozik a tudományos kommunikáció is, amely alapvetően függ a tudományos eredmények gyors és széles körű disszeminációjától. Feltartóztathatatlan fejlemény a globális tudományos élet számára egy lingua franca használata, amely jelenleg az angol nyelv. Emellett azonban nem mondhatunk le arról, hogy a tudományos eredmények közvetítése, illetve a képzés az anyanyelven történjék.

A Nyelvtudományi Kutatóközpont, illetve jogelődje, az MTA Nyelvtudományi Intézete (NYTI) nagy tapasztalattal rendelkezik mind a szövegkorpuszok építése, mind pedig a dokumentumok digitálisan feldolgozása terén. A korpuszépítések közül talán a legismertebb a Magyar Nemzeti Szövegtár (VÁRADI 2002, VÁRADI–ORAVECZ 2014). A most bemutatandó kutatásnál azonban a klasszikus korpuszépítés legfontosabb is-

* Az MTA Tudomány a Magyar Nyelvért Nemzeti Program II. alprogramjának bemutatása.

DOI: https://doi.org/10.18349/MagyarNyelv.2023.4.478

mérvein túl (reprezentativitás, kiegyensúlyozottság) olyan feladatok kerülnek a középpontba, mint például az egyes szövegekre jellemző metaadatok kinyerése és rendszerezése, a megfelelő struktúrák kidolgozása és felépítése (Laki et al. 2022). A korpuszépítéshez felhasznált szövegek tekintetében egyre nagyobb figyelmet kapnak a közgyűjteményekben található anyagok. A hazai szöveges dokumentumtárak közül az MTA Könyvtárának repozitóriuma, a REAL előkelő helyet foglal el mind a tartalmazott szövegek mennyiségében, mind a gyűjtőkör (tudományos jellegű, lektorált szövegek) tekintetében. Ráadásul az MTA Könyvtár és Információs Központ (MTA KIK) Magyarországon az elsők között látott hozzá repozitórium építéséhez, és az anyagok gyűjtésében, gondozásában is nagy tapasztalattal rendelkezik (Holl 2013). Mivel az MTA NYTI részt vett a MATRICA projektben, amely hivatkozások kinyerését tűzte ki célul tudományos folyóiratcikkekből (Váradi et al. 2014), továbbá ezen a téren több nemzetközi együttműködésben is (Confederation of Open Access Repositories, European Open Science Cloud), más részről az MTA KIK kiváló tapasztalatokat szerzett az egyedi azonosítók alkalmazásában és a nemzeti bibliográfiai adatbázisok fejlesztésében (Holl 2022a, Sīle et al. 2018), könnyen adódott az együttműködés gondolata.

2. A kutatás fő célkitűzései. A kutatás célja kettős: egyrészt segíteni a tudományos kommunikációt gépitanulás-alapú módszerekkel, másrészt komoly támogatást nyújtani a magyar nyelvű tudományosság számára a magyar nyelvű tudományos közlések feldolgozásával és magasabb szintű elérhetővé tételével. A konzorcium olyan új megközelítésű eljárások fejlesztését végzi, melyek eddig egyáltalán nem, vagy tudományos szolgáltatásokban még nem valósultak meg. Elsősorban az ún. born digital, tehát digitális anyagként született lektorált tartalmak kinyerése a cél, amelyek jól használhatók a gépi tanuló rendszerek tanítóanyagaként is. A kifejlesztett eszközök az MTA KIK adatbázisainak javítására, gazdagítására szolgálnak a szövegek automatikus szemantikus szegmentálásával, a szakterület megállapításával, a metaadatok kinyerésével és javításával, sőt, a szöveghibák javításával és a szövegbeli szakkifejezések kinyerésével. Ezáltal a kutatás egy széles értelemben vett szövegbányászatot (Holl 2015) valósít meg, ahol az új nyelvtechnológiai eszközök lehetővé teszik nemcsak a REAL repozitórium tartalmának, hanem a felhasználók számára nyújtott repozitóriumi szolgáltatásoknak is a javítását (Holl 2022b). Az MTA KIK-nél tárolt anyagok ezáltal a korábbinál hatékonyabb módon válnak kutathatóvá, például azáltal, hogy az MTMT hazai folyóiratokban megjelent, az adatbázisból eddig hiányzó hivatkozásokkal is bővül.

3. A kutatás menete. A REAL repozitórium szöveges tartalmainak feldolgozása többféle technológiai kihívást jelent. Ezek közül az első a leggyakrabban pdf-formátumban meglévő tartalmak olyan nyers reprezentációra alakítása, amely alapul szolgálhat a további számítógépes feldolgozás számára. További kihívást jelent a szövegek szűrése, szegmentálása. A szövegek konverziójának eredményeképpen létrejött korpusz a munkálatok első közös mérföldköve, amelyet a könyvtártani és nyelvtechnológiai célú szövegbányászat követ. Az NYTK jelentős tapasztalattal rendelkezik olyan, a feldolgozáshoz szükséges legfontosabb nyelvtechnológiai eszközök létrehozásában, mint a névelem-felismerő vagy az automatikus címkéző.

Könyvtártani szempontból kiemelkedő fontosságú a szövegek szerkezetének, szemantikai elemeinek automatikus felismerése (cím, szerzők, affiliációk, hivatkozások, köszönetnyilvánítások). További feladat a hagyományosan tulajdonneveknek nevezett szerkezeteknél bonyolultabb névkifejezések (személynevek, földrajzi nevek, szervezetek nevei, egyedi azonosítók (PID), pályázati azonosítók, webcímek stb.) kinyerése. A PID-ek ugyanakkor viszonylag könnyen felismerhetők, és kinyerésük lehetővé teszi a szövegek kapcsolati hálózatának feltérképezését (mind társszerzői-idézői hálózatok, mind intézményi és támogatási hálózatok tekintetében). A tulajdonnevek kinyerése – névterek felhasználásával – további szemantikus keresési lehetőségek alapjául szolgálhat.

A projekt során magyar nyelvű nyomtatott dokumentumokban található szkennelt, majd OCR-ezett, illetve „born digital" szövegeket dolgozunk fel. A pdf-anyagok szöveges formátumra (.txt) konvertálását és tisztítását követően egy automatikus nyelvi szűrés végzi a REAL gyűjteményben megtalálható idegen nyelvű szövegek eltávolítását, majd az e-magyar rendszer (VÁRADI et al. 2017) segítségével mind a digitalizált, mind a digitálisan létrejött szövegek nyelvi annotációja következik. Ezt követően a repozitórium fejlesztését segítő tanítóanyagok létrehozása történik meg a szövegelemek (köszönetnyilvánítás, bibliográfia, absztrakt, cím, szerző, kulcsszó stb.) gépi felismerésével és a bibliográfiai adatok elemekre bontásával. Bármennyire is vonzó a tanítóanyagok teljesen automatikus létrehozása, a tanítóanyagok pontossá tételében az emberi átolvasás és a kézi címkejavítások nem kerülhetők el.

Statisztikai módszereket és szabályokat is alkalmazunk az egyes szaknyelvekre jellemző szakkifejezések, azaz terminusok automatikus kinyeréséhez. Ennek segítségével valósul meg a Magyar Egységes Ontológiába, illetve a könyvtárakból jól ismert ETO-hierarchiába rendezés. A fejlesztések eredményei integrálódnak a meglévő webes szolgáltatásokba, illetve szükség esetén új webes felület is kialakításra kerül az MTA könyvtárának repozitóriuma számára. A szövegbányászat eredményeivel kiegészített metaadat-készlet a repozitórium keresésének hatékonyabbá tételét eredményezi.

4. A kutatás várható eredményei és további tervek. A nyelvtechnológia abban tudja tehát segíteni a repozitórium anyagában való tájékozódást, hogy a tudományos publikációk óriási tömegének tartalmát teszi könnyen kereshetővé a pdf-alakból automatikusan kinyert metaadatok segítségével. A szövegek nyelvtechnológiai feldolgozása több szinten halad: a szerzők és affiliációik, a hivatkozások, a köszönetnyilvánítások, az igénybe vett infrastruktúrák megnevezésétől a szakterminológia, a névkifejezések feldolgozásán át a komplexebb szemantikai tartalmak kinyeréséig.

Az együttműködés kölcsönösen hasznos. A REAL magyar nyelvű publikációiból olyan korpusz épül a magyar nyelv neurális hálós modelljeihez, amely a normatív nyelvhasználatot, a gondozott szövegeket reprezentálja, aminek kiemelt értéke van napjaink nyelvtechnológiai kutatásainak világában, hiszen az ismert eljárások túlnyomó része az internetről szerzett, azaz a hivatalos normát nem feltétlenül követő szövegeken alapul.

Az együttműködés egyaránt szolgálja a magyar nyelv és a magyar nyelvű tudományosság digitális fenntarthatóságát. A jelzett kutatás támogatásához fontos kiemelni a magyar nyelvű tudományos terminológia központi szerepét, hiszen a Nyelvtudományi Kutatóközpontban folyik egy Nemzeti Terminológiai Központ kialakítása, melynek

hosszú távú célja a tudomány különböző területein a magyar nyelvvel kapcsolatos tevékenységek koordinálása.

Az MTA KIK fontos adatbázisokat szolgáltat a hazai tudományos közösség számára: a REAL-ban tárolt tudományos szövegek nem csupán a nyelvtechnológiai eszközök fejlesztésének nyersanyagát képezik, hanem a fejlesztett eszközök lehetővé teszik a REAL és az MTMT szolgáltatásának javítását is. A fejlesztendő informatikai eszközök olyan mértékű adatjavítási és -gazdagítási potenciállal rendelkeznek, amelyet könyvtáros munkaerők alkalmazásával a gyakorlatban nem lehetne elvégezni.

Hivatkozott irodalom

Holl András 2013. Az Akadémiai Könyvtár repozitóriuma, a REAL bővítése, gyarapítása, fejlesztése. *Tudományos és Műszaki Tájékoztatás* 60: 198–199.

Holl András 2015. Szövegbányászat, adatbányászat, ismeretfeltárás. Új lehetőségek a tudományos kommunikációban. *Magyar Tudomány* 176: 680–685.

Holl András 2022a. Repozitóriumok – különleges terület a könyvtárak világában. *Tudományos és Műszaki Tájékoztatás* 69: 358–365. http://doi.org/10.3311/tmt.13180

Holl András 2022b. A hazai tudományos eredmények láthatóvá tétele, kiaknázása és megőrzése modern eszközökkel. *Magyar Tudomány* 183: 69–78. http://doi.org/10.1556/2065.183.2022.1.6

Jelencsik-Mátyus, Kinga et al. 2022. Report on the Hungarian Language. *European Language Equality* D1.18. http://doi.org/10.1007/978-3-031-28819-7_20

Kornai, András 2013. Digital Language Death. *PLoS ONE* 8: e77056. https://doi.org/10.1371/journal.pone.0077056

Laki, László János et al. 2022. OCR-hibák javítása neurális technológiák segítségével. In: Berend Gábor – Gosztolya Gábor – Vincze Veronika szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem Informatikai Intézet, Szeged. 417–430.

Sīle, Linda et al. 2018. Comprehensiveness of national bibliographic databases for social sciences and humanities: Findings from a European survey. *Research Evaluation* 27: 310–322. https://doi.org/10.1093/reseval/rvy016

Váradi Tamás 2002. The Hungarian National Corpus. In: Rodríguez, Manuel González – Suarez–Araujo, Carmen Paz eds., *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*. European Language Resources Association, Paris. 385–389.

Váradi Tamás et al. 2014. Magyar társadalomtudományi citációs adatbázis: a MATRICA projekt eredményei. In: Tanács Attila – Varga Viktor – Vincze Veronika szerk., *X. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 269–276.

Váradi Tamás et al. 2017. Az e-magyar digitális nyelvfeldolgozó rendszer. In: Vincze Veronika szerk,. *XIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 49–60.

Váradi Tamás – Oravecz Csaba 2014. A Magyar Nemzeti Szövegtár egymilliárd szavas új változata. *Magyar Tudomány* 175: 1054–1061.

The digital support of the Hungarian language in support of Hungarian science

The Repository of the Library and Information Centre of the Hungarian Academy of Sciences (REAL) is an important secondary (archived) source of scientific literature in Hungarian. While in the past this collection served individual researchers' document needs in accordance with traditional library functionality, here the text layers of documents are treated as a corpus of text. Linguistic tools are used to explore and mine the corpus in a broad sense, including the extraction of references to literature and recognition of various named entities. The project will improve the quality of the text by the identification of possible textual errors and enrich the metadata of the documents. The objective of the project is to improve both repository services and data quality, enabling the development of value-added services for the research community.

Keywords: repositories, text corpora, automated annotation.

Prószéky Gábor – Váradi Tamás
HUN-REN Nyelvtudományi Kutatóközpont

Holl András
MTA Könyvtár és Információs Központ

A magyar nyelv digitális fenntarthatóságának támogatása*

1. A kutatás előzményei

A Nyelvtudományi Kutatóközpont (NYTK) most ismertetendő munkálatainak a célja annak a Magyar Tudományos Akadémia (MTA) alapításakor vállalt küldetésnek a biztosítása, miszerint anyanyelvünk ma a digitális térben is betölthesse méltó szerepét. A nyelvek digitális támogatását célzó nemzetközi élvonalbeli kutatások a legnagyobb beszélőszámmal rendelkező nyelvekre, elsősorban az angolra összpontosítanak, és kevés figyelmet szentelnek a kisebb piacot jelentő nyelvekre. A magyar nyelv technológiai támogatása nemzeti ügy, amelyhez elengedhetetlen az élvonalbeli technológiai eszközök, valamint a létrehozásukhoz és működtetésükhöz szükséges írott és hangzó adatbázisok elkészítése a magyarra is. Az ismertetendő munkálatok keretében végzett kutatások a normatív magyar nyelvre fókuszálnak, tehát nem az interneten megjelenő szövegekre válogatás nélkül, hanem a gondozott, szerkesztőségi kontrollon átmenőkre, valamint a magyar nyelv határon kívüli és belüli változataira, továbbá a rokon uráli nyelvekre.

A legtöbb európai nyelv digitális infrastruktúrájának egyik központi pillére a nemzeti korpusz, amely az adott nyelv hiteles, reprezentatív mintáját jelenti. A Magyar Nemzeti Szövegtár (MNSZ.) jelenleg egy több mint egymilliárd szavas magyar elemzett korpusz,

* Az MTA Tudomány a Magyar Nyelvért Nemzeti Program III. alprogramjának bemutatása.

DOI: https://doi.org/10.18349/MagyarNyelv.2023.4.482