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

amely hat stílusrétegből (sajtó, tudományos, szépirodalmi, hivatalos, személyes és beszélt nyelvi) tartalmaz szövegeket. Ezeken belül pedig az anyag öt regionális nyelvváltozatra oszlik, bemutatva ezzel a magyarországi mellett a határon túli magyar nyelvhasználatot is (VÁRADI 2002; VÁRADI–ORAVECZ 2014). Az adatbázist elsősorban a magyar nyelvre vonatkozó korpuszalapú és korpuszvezérelt nyelvészeti kutatások támogatására használják, nemcsak a nyelvészet területén, hanem a bölcsészet- és társadalomtudományokban is számos vonatkozásban a pszicholingvisztikától a diskurzuselemzésig (lásd pl. SASS 2005). A szöveg- és tartalomelemzés területén ez a legtöbbet használt magyar nyelvi forrás. A művelt nagyközönség körében is igen népszerű, a regisztrált használók száma meghaladja a tizenötezret. Az új lehetőségek, mint például az újabb, jobb minőségű nyelvi elemzők (OROSZ et al. 2022) és az újabb kihívások, például a nyelvi modellek előállításához szükséges nagyméretű adatbázisok (NEMESKEY 2020) azonban szükségessé teszik a korpusz bővítését és fejlesztését.

A *helyesírás* a nyelvi norma intézményesülésének egyik központi tényezője, amely az egyes nyelvi kultúrákban eltérő mértékben ugyan, de általában explicit módon szabályozó szerepet játszik. A magyar helyesírási norma egy absztrakt helyesírási szabályzatban (MTA 2015), valamint ahhoz jelentős számú konkrét példát tartalmazó helyesírási szótárban (MTA 2016) testesül meg. A norma közvetítése és betartása a történelem folyamán mindig a művelt elit feladata volt. A digitális társadalom világában ez a hagyományos kép alapvetően megváltozott: a digitális tér az internet jóvoltából drasztikusan kitárult, és olyan emberek tömegei jutottak szóhoz a nyilvánosság előtt, akiknek azelőtt esélyük sem volt erre. A digitális kommunikációs térben tömegesen jelennek meg nem normakövető szövegek, amelyek viszont olykor elbizonytalanítják az olvasókat, így sokan a korábbiaknál nagyobb mértékben rá vannak utalva a normát jól ismerők tanácsára. A digitális eszközök ma már a korábbiaknál teljesebb és rugalmasabb megoldást kínálnak az elvont és száraz szabályzat és a példákat felsoroló (tehát óhatatlanul hiányos) papír szótárakkal szemben. A 2013-ban az MTA segítségével útjára bocsátott Helyesírási tanácsadó portál (https://helyesiras.mta.hu) ezt a társadalmi igényt hivatott szolgálni, korszerű nyelvtechnológiai módszerek felhasználásával (MIHÁLTZ et al. 2012). A portál azonban megérett az újításra mind a szoftverplatform, mind pedig a módszertan és az ügyfélközpontúság terén.

A magyar nyelv nagyszótárához készült archiválís cédulagyűjtemény első darabjainak összeválogatása már a 19. század végén megkezdődött, és bár a rendszerezésükkel még sok munka adódott, a számuk az 1950-es évekre meghaladta a négymilliót[1] (LIPP–SIMON 2021). A túlnyomórészt A/6-os méretű és kézzel írt lapokból álló kollekció annak a „közösségi gyűjtés”-nek az eredményeként jöhetett létre, amely a Magyar Nyelvőr című folyóiratban az 1890-es években publikált felhívásokat és a Magyar Tudományos Akadémia által 1899-ben közzétett „Utasítás”-t követően indult el, és tartott közel öt évtizeden át. A cédulákat alapítása óta a Nyelvtudományi Kutatóközpont (illetve annak jogelődje) erre kialakított termében tárolták. A szótár munkatársai a cédulaanyag bővítését, az adatgyűjtést az 1950-es években is folytatták. Ezzel párhuzamosan folyt egy olyan gyűjtőmunka is, amelynek célja egy adott szöveg minden egyes szóelőfordulásának rögzítése volt: ilyen módon a 20. századi nagy magyar költők közül például Ady Endre és József

[1] Címszójegyzéke megtalálható: https://nagyszotar.nytud.hu/slips.html.

Attila számos versének, továbbá az alkotmány teljes szövegének az összes szóadata is dokumentálva lett. A cédulaanyagot 1974 és 1995 között rendezték betűrendbe, majd megkezdődött a legépelése (KISS 2004), de teljes körű digitális használatához még további fejlesztések szükségesek.

A magyar rokon nyelvei közül a hanti és a manysi képviseli az obi-ugor ágat. A magyar nyelv digitális jelenléte is támogatásra szorul, de a kis rokon nyelvek revitalizációjára a digitális térben talán még nagyobb szükség van (JELENCSIK-MÁTYUS et al. 2022). Ehhez szükséges lenne – a magyar mintájára – egy elemzett digitális korpuszt létrehozni, amely hanti és manysi anyagokat tartalmaz, lehetőleg minél több annotációval. A kétezres évek elején ugyan már készült egy annotált korpusz a hanti és a manysi nyelvek különböző nyelvjárásainak feldolgozásához, de ezek helyszíni beszélt nyelvi gyűjtések hanganyagainak átírásával jöttek létre (FEJES–NOVÁK 2010). Emellett (szinjai és szurguti) hanti anyagokat találunk az uráli nyelvek mondattanának vizsgálatára létrejött korpuszban is[2] (SIMON 2017). Mindkét esetben egy uráli nyelvekre létrehozott elemzőt[3] használtak a szövegek morfológiai annotálására (FEJES–NOVÁK 2010). Jelen projektben is ezeket az eszközöket tervezzük használni, azonban feltétlenül szükséges annak átalakítása a feldolgozandó szövegek tulajdonságait szem előtt tartva – ezzel segítve a további nyelvészeti kutatásokat. A jelen programban a korpusz alapját (a magyar korpuszhoz hasonlóan) írott, modern szövegek képezik majd, melyek lehetőséget adnak a nyelvek jelenlegi állapotának dokumentálására és megőrzésére.

2. A munkálatok célkitűzései

A jelen munkálatok tehát a fent ismertetett négy téma mentén négy olyan alprojekt keretében végzett kutatásokra és fejlesztésekre irányulnak, amelyekben a Nyelvtudományi Kutatóközpont mint az MTA Kiváló Kutatóhelye elismerten élvonalbeli és egyedi szerepet tölt be. Ezek az alábbiak: 1. Magyar Nemzeti Szövegtár 3.0; 2. Helyesírási tanácsadó portál 2.0; 3. A magyar nyelv nagyszótára archivális cédulagyűjteményének digitalizációja; 4. Obi-ugor nyelvek elemzett korpusza és tudásbázisa.

A NYTK eddig is küldetésének tekintette a magyar nyelv és nyelvhasználat korszerű technológiai eszközökkel történő támogatását. A Kutatóközpontban működő Nyelvtechnológiai Kutatócsoport az MTA támogatásával több olyan nyelvi erőforrást és digitális eszközt hozott létre, amelyek népszerűnek bizonyultak a szakmai kutatók és az érdeklődő nagyközönség körében egyaránt. Ezek között a most ismertetendő projekt keretében a Magyar Nemzeti Szövegtár, valamint a helyesiras.mta.hu címen működő helyesírási tanácsadó portál megújítását kívánjuk elvégezni.

A fent bemutatott két, korábban a NYTK-ban elkezdett munka mellett a projekt fontos részét képezi az archivális cédulagyűjtemény anyagának teljes digitalizálása, amely elsősorban kulturálisörökség-védelmet jelent. Emellett azonban lehetőség nyílna ennek segítségével további lexikológiai kutatások folytatására is. Az internetes publikálást követően az archivális cédulagyűjtemény a laikus és szakmai közönség számára is elérhetővé válik, nem beszélve arról, hogy a nagyszótári projekten dolgozó lexikográfusok is könnyebben hozzáférnek az anyaghoz, mint eddig.

[2] http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html

[3] https://morphologic.hu/urali

A negyedik alprojekt keretében nyelvek átfogó leírása történik meg (Bakró–Laakso–Skribnik 2022), amelyben a NYTK munkatársai mutatják be az obi-ugor nyelveket (lásd például Sipos 2022), ezzel biztos alapokat adva a nyelvi elemzők fejlesztéséhez.

Mind a négy alprojektet a Nyelvtudományi Kutatóközpont kutatóinak irányításával működő interdiszciplináris csapat végzi, amelyben a nyelvész és nyelvtechnológus szakértők mellett informatikusok, szoftverfejlesztők, webfejlesztők dolgoznak együtt. A munkálatok tervezett ideje 48 hónap.

3. Az egyes részprojektek

3.1. Magyar Nemzeti Szövegtár 3.0. A korpusz első változatában még csak 187,6 millió szavas MNSZ. második kiadása 2014-ben jelent meg, és több mint 1 milliárd szóra bővült. Bár az új változatot az akkor elérhető legjobb minőségű számítógépes nyelvészeti elemzőprogram segítségével dolgoztuk fel, azóta sokkal hatékonyabb eszközök és gazdagabb adatkészletek állnak rendelkezésünkre. Ezekre alapozva a 3.0-s verzióhoz az alábbi fejlesztéseket tervezzük:

a) A korpusz mérete a jelenlegi tízszeresére, azaz 10 milliárd szóra nő.

b) Bevezetjük a forrásszövegek szövegtípusok szerinti osztályozását.

c) A MNSZ. jelenlegi változatának elkészülte óta létrehozott nyelvi elemzőrendszer, az e-magyar integrálásával új nyelvi annotációt kap, amely nemcsak pontosabb és mélyebb nyelvi kereshetőséget jelent, hanem a vizsgálható jelenségek szélesebb körére (névkifejezések, terminológia, mondatszerkezet) is kiterjed.

d) Egy ilyen nagy méretű, több szinten annotált szövegkorpusz komplex lekérdezéséhez speciális eszközre van szükség, ezért az új adatbázist a tervek szerint a KorAP platformba illesztjük be, melyet a Leibniz Institut für Deutsche Sprache fejlesztett ki (https://www.ids-mannheim.de/en/s/corpus-linguistics/projects/korap).

3.2. Helyesírási tanácsadó portál 2.0. A https://helyesiras.mta.hu portál hét különböző nyelvtechnológiai eszköz segítségével támogatja a felhasználókat abban, hogy megtalálják a megoldást a helyesírási problémáikra. A megfelelő eszközt használva például választ kaphatunk arra, hogy az adott szavakat egybe vagy külön kell-e írni, vagy megismerhetjük egy-egy szó, ill. kifejezés helyesen írt alakját/alakjait, és a számok, valamint a dátumok helyesírását is ellenőrizhetjük. Mindemellett a portál ismerteti az érvényben lévő helyesírási szabályzat magyarázatait is, tehát segít megérteni és megtanulni a szabályt, hogy a későbbiekben önállóan is tudjuk alkalmazni. A jól használható alkalmazások, és az, hogy a legtöbb egyszerű helyesírási kérdésre választ kaphatunk, hamar népszerűvé tették a weboldalt. A portálhasználatot elemző legfrissebb statisztikák napi mintegy 70.000 oldallátogatást mutatnak, összesítve ez az erősebb hónapokban túllépi az 1.700.000-t is.

A nagy népszerűségnek örvendő Helyesírási tanácsadó portál mára több szempontból megérett az – elsősorban technológiai – újításra, mert a készítésekor modernnek számító programnyelv mára elavulttá vált, és a szoftverplatform, amelyen fut, már nem eléggé biztonságos. A Helyesírási tanácsadó portál 2.0 verziójához az alábbi fejlesztéseket végezzük el:

a) A kódbázis lecserélése Python 3 nyelvű változatra.

b) A háttéradatbázisok frissítése.

c) Kontextusérzékeny tanácsadás mesterséges intelligenciát használó interaktív technológia segítségével.

d) A portál internetes felületének frissítése.

3.3. A magyar nyelv nagyszótára archiválisc édulagyűjteményének digitalizációja. Az archiválisc édulagyűjtemény anyagának egy része az elmúlt évtized folyamán már digitalizálásra került. A hátralévő 822 doboznyi, mintegy 3.400.000 darab A/6-os cédula digitalizálása reprográfiai szakértők, informatikai munkatársak és annotátorok közreműködésével, a Lexikai tudásreprezentáció kutatócsoport szakmai irányítása mellett valósul meg. Mivel a túlnyomórészt kézzel írt szövegeket tartalmazó cédulákról készített képek további feldolgozása is tervbe van véve, az előállítandó anyaggal szembeni minőségi követelményeket a Magyar Nemzeti Levéltárral folytatott szakmai konzultáció során állapítjuk meg. A digitalizálási munkát a kutatóközpont tulajdonában lévő professzionális dokumentumszkennerekkel végezzük. Az alprojekt a következő munkaszakaszokra oszlik: előkészítés, digitalizálás (valamint az elkészült anyagok rendszerezett tárolása), annotálás és publikálás.

3.4. Obi-ugor korpusz és tudástár. A negyedik alprojekt az obi-ugor nyelvek, a hanti és a manysi megőrzését célzó annotált nyelvi korpusz létrehozása. A hanti az uráli nyelvcsalád ugor ágához tartozó agglutináló nyelv, melynek morfológiája és szintaxisa sok tekintetben hasonlít a magyaréhoz. A nyelv maga veszélyeztetett, a generációk közötti nyelvátadás nincs biztosítva. A manysi a hanti nyelv legközelebbi rokona. Tipológiai, morfológiai és szintaktikai szempontból ugyanúgy jellemezhető, mint a hanti. A manysi változatai közül ma már gyakorlatilag csak az északit beszélik, és a kiadványok is ebben a nyelvjárásban jelennek meg.

A korpuszépítés alapjául online megjelenő újságok szövegei fognak szolgálni. A Ленин пант хуват (Lenin pant xuwat, azaz 'A Lenini Úton') című újság az 1950-es évektől kéthetente jelenik meg. Kezdetben hanti és manysi nyelvű cikkek vegyesen szerepeltek benne, melyek oroszul is olvashatók voltak, azonban ezekből 1990 körül létrejött egy-egy külön hanti és külön manysi nyelvű újság. A továbbra is kéthetente megjelenő hanti nyelvű kiadvány elnevezése Хӑнты ясӑӈ (Xanti jasang, azaz 'Hanti Szó') lett. Az újságnak a honlapján archívuma is van, és a tartalmak mind hanti, mind pedig orosz nyelven egyaránt olvashatók. A manysi nyelvű újság neve Luima Seripos*,* és szintén három évtizede jelenik meg kétheti rendszerességgel az interneten. Jelenleg a weben a 2012 utáni számok érhetők el.

Az újságírás műfajai közül mindkettőben elsősorban a hír, a riport és az interjú fordul elő leggyakrabban. Ezek a kiadványok kezdetben mindössze néhány oldalasak voltak, mára azonban 16 oldalas, színes képekkel ellátott újságok lettek. Napjainkban mindkét kiadvány cirill alapú mellékjeles karaktereket alkalmaz, míg korábban a cirill betűket nem, vagy csak kis mértékben egészítették ki diakritikus jelek.

A korpuszon morfológiai elemzést és névelem-felismerést végzünk, egy korábban elkészült obi-ugor morfológiai elemzővel és az e-magyar elemzőlánc névelem-felismerő eszközével. Az alprojekten belül elvégzendő feladatok a következők: forrásanyaggyűjtés és –előfeldolgozás, morfológiai elemzés, névelem-felismerés, a korpusz beillesztése egy korpuszlekérdező rendszerbe.

4. A kutatás várható eredményeiről

4.1. Magyar Nemzeti Szövegtár 3.0. A mesterséges intelligencia kutatásában és fejlesztésében kulcsszerepet játszik a nyelvtechnológia, és különösen fontos az elérhető korpuszok mérete és minősége. A projekt eredményeként létrejövő 10 milliárd szavas, nyelvi elemzéssel ellátott korpusz feldolgozásához egy olyan lekérdezőrendszert használunk majd, amely nemcsak a jelenleginél modernebb lehetőségeket biztosít a használók számára, hanem ezáltal a MNSZ. 3.0 az európai nemzeti korpuszok hálózatának is részévé válik.

4.2. Helyesírási tanácsadó portál 2.0. A weboldal alapjául szolgáló platform elkerülhetetlen technológiai váltását össze kívánjuk kötni a portál szolgáltatásait érintő tartalmi és módszertani újításokkal. A kilenc évnyi működtetés alatt összegyűlt tapasztalatok kiértékelése, a felhasználói visszajelzések rendszeres áttekintése szilárd alapul szolgál arra, hogy tartalmában, terjedelmében és a felhasználói élmény szempontjából egyaránt megújítsuk a Helyesírási tanácsadó portált. A tervezett 2.0-s változat egyik alapkoncepciója az, hogy a portál még hatékonyabban, még testreszabottan szolgálja a felhasználók igényeit. Ennek egyik kulcseleme annak a nyelvi kontextusnak a pontosabb ismerete, amelyben a kérdéses alak előfordul. Ennek ismerete nélkül a jelenlegi rendszer gyakran kénytelen meglehetősen általános választ adni, azaz olyat, amely mindkét vagy mindhárom kérdéses változat használatát is jóváhagyja. Ez korrekt, de nyilvánvalóan nem feltétlenül hatékony megoldás, és olykor frusztrációt is okozhat a rendszer használóinak. A fejlesztésnek ez a része jelenti a legizgalmasabb kihívást, melyet mesterséges intelligenciát használó interaktív technológiával tervezünk megoldani.

4.3. A magyar nyelv nagyszótára archivális cédulagyűjteményének digitalizációja. Az archivális cédulagyűjtemény anyagának digitalizálása kulturális örökségünk egy unikális „entitásának" megőrzése mellett a gyűjtemény feldolgozhatóságához is hozzájárul. A kézzel írott cédulák képein lefuttatott karakterfelismertetéssel olyan kereshető szövegű adatbázist, korpuszt kapunk majd eredményül, amely lehetővé teszi a gyűjtemény információtartalmának feltárását. Mindez A magyar nyelv nagyszótára munkálataiban is jelentős gyorsulást fog eredményezni, mivel a szótár példaanyagának egyik fontos forrása ez a gyűjtemény. Az elmúlt száz évben csak papíron elérhető anyag a digitalizálást és az interneten való publikálást követően az érdeklődök széles köre számára válik elérhetővé és egyben kutathatóvá.

4.4. Obi-ugor korpusz és tudástár. A kevés beszélővel rendelkező nyelvek digitális térben történő revitalizációjára nagy szükség van a nyelv fennmaradásához, és ahhoz, hogy az egyre terjedő digitális szolgáltatások a saját nyelvükön valósulhassanak meg. Minden korpusz, amely valamely kis nyelv anyagait tartalmazza, egy újabb lépés és lehetőség afelé, hogy ne vegye át az angol nyelv a szolgáltatásokat. A projektben az anyaggyűjtésen túl a hanti és manysi nyelv morfológiai elemzése, valamint a névelemfelismerés is szerepel, amely jelentősen megkönnyíti a magyar két közeli rokon nyelvének digitális támogatását.

5. További tervek

A vázolt fejlesztések mind a magyar, mind a közeli rokon nyelvek digitális fenntarthatóságának támogatását célozzák.

A jelenleg létező, szerkesztett szövegeket tartalmazó MNSZ. 2. újabb nagyságrenddel való növelése, valamint annak a KorAP platformba való beillesztése jelentősen megnöveli mind a korpusz használóinak lehetőségeit, mind pedig a korpusz láthatóságát azzal, hogy

az európai nemzeti korpuszok hálózatának részévé válik. Egyben egy ekkora adatgyűjtemény jelentős alapját képezheti további nyelvészeti kutatásoknak és fejlesztéseknek is.

Bár digitális eszközeink képesek a segítségünkre lenni az egyszerűbb helyesírási kérdésekben, az összetettebb jelenségek, például az egybe- vagy különírás témájában való segítségnyújtás komplexebb hátteret kíván, amit a felújított helyesírási portál már az ügyfélelégedettség szem előtt tartásával igyekszik megvalósítani.

A magyar nyelv nagyszótára archiválisc édulagyűjteménye digitalizációjának kapcsán a kutatás jelentősége a cédulákon található kézírásos szövegeknek a – Magyar Nemzeti Levéltárral való együttműködés keretében történő – felismertetése, azaz a cédulákra kézzel kiírt idézeteknek a kereshető szövegadatbázissá alakításának sikerességében, valamint az egyes képfájloknak a cédulákról kinyert címszó alapján eszközölt gépi annotálása eredményességének vizsgálatában ragadható meg.

A kisebb uráli nyelvek, a magyar nyelv legközelebbi rokonai, a hanti és a manysi veszélyeztetettek, ezért dokumentálásuk és digitális támogatásuk kiemelkedő fontosságú feladat. Az annotált korpusz létrejötte lényegében egy kezdő lépés lehet számos kutatás és fejlesztés számára, amelyek hozzájárulhatnak e nyelvek revitalizációjához.

Hivatkozott irodalom

Bakró-Nagy, Marianne – Laakso, Johanna – Skribnik, Elena eds. 2022. *The Oxford Guide to the Uralic Languages*. Oxford University Press, Oxford. http://doi.org/10.1093/oso/9780198767664.001.0001

Fejes László – Novák Attila 2010. Obi-ugor morfológiai elemzők és korpuszok. In: Tanács Attila – Vincze Veronika szerk., *VII. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 284–291.

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

Kiss Gábor 2004. A Nagyszótár címszójegyzékéről és az archivális cédulagyűjtemény nagyságáról, gazdagságáról. In: Fóris Ágota – Pálfy Miklós szerk., *A lexikográfia Magyarországon*. Tinta Könyvkiadó, Budapest. 39–52.

Lipp, Veronika – Simon, László 2021. Towards a new monolingual Hungarian explanatory dictionary: overview of the hungarian explanatory dictionaries. *Studia Lexicographica* 15/29: 83–96. http://doi.org/10.33604/sl.15.29.4

Mihálcz Márton et al. 2012. Helyesírás.hu – Nyelvtechnológiai megoldások automatikus helyesírási tanácsadó rendszerben. In: Tanács Attila – Vincze Veronika szerk., *IX. Magyar Számítógépes Nyelvészeti Konferencia.* JATEPress, Szeged. 135–148.

Magyar Tudományos Akadémia közread. 2015. *A magyar helyesírás szabályai.* Tizenkettedik kiadás. Akadémiai Kiadó, Budapest.

Magyar Tudományos Akadémia közread. 2016. *Magyar helyesírási szótár: a magyar helyesírás szabályai 12. kiadása szerint.* Akadémiai Kiadó, Budapest.

Nemeskey, Dávid Márk 2020. *Natural Language Processing methods for Language Modeling*. Eötvös Loránd Tudományegyetem, Budapest. Doktori disszertáció. http://doi.org/10.15476/ELTE.2020.066

Orosz, György et al. 2022. HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In: Berend Gábor – Gosztolya Gábor – Vincze Veronika szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem Informatikai Intézet, Szeged. 59–73.

Sass Bálint 2005. Vonzatkeretek a Magyar Nemzeti Szövegtárban. In: Alexin Zoltán – Csendes Dóra szerk., *III. Magyar Számítógépes Nyelvészeti Konferencia*. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 257–264.

Simon Eszter 2017. Négy hatás alatt álló nyelv – Korpuszépítés kis uráli nyelvekre. In: Vincze Veronika szerk., *XIII. Magyar Számítógépes Nyelvészeti Konferencia.* Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged. 263–274.

Sipos, Mária 2022. North Khanty. In: Bakró-Nagy, Marianne – Laakso, Johanna – Skribnik, Elena eds., *The Oxford Guide to the Uralic Languages*. Oxford University Press, Oxford. 582–607. https://doi.org/10.1093/oso/9780198767664.003.0031

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

For the digital sustainability of the Hungarian language

The project follows the founding mission of the Hungarian Academy of Sciences to ensure that Hungarian is given a worthy role in the digital space. International research focuses mainly on English, with less attention paid to smaller languages like Hungarian. (1) The Hungarian National Corpus (MNSz) consists of more than one billion words, and is highly used in linguistic research. It is composed of six stylistic layers and five regional language varieties. The corpus is primarily used to support corpus-based and corpus-driven linguistic research on Hungarian, not only in linguistics but also in many fields of the humanities and social sciences. However, new possibilities, such as newer, higher-quality language parsers or large databases to produce large language models, have made it necessary to expand and improve the corpus. (2) Spelling control is a key element of the linguistic norm and is becoming increasingly important in the digital space. The Spelling Advisory Portal, supported by the Hungarian Academy of Sciences, meets this demand with state-of-the-art technology, but needs upgrading in terms of software platform, methodology and customer focus. (3) The collection of more than four million dictionary cards belonging to the Great Dictionary of the Hungarian Language was created at the end of the 19th century. The cataloguing and digitization of the dictionary cards is ongoing and, although the construction of the collection started almost twenty years ago, further development is needed to ensure its full digital use. (4) In order to support the digital presence of the closest relatives of Hungarian: Khanty and Mansi, there is a need to create an analyzed digital corpus based on written modern texts that provide the opportunity to document and preserve the current state of the Ob-Ugric languages. In summary, these studies will contribute to the development of Hungarian in the digital space and cover a wide range of linguistic research from normative language to orthography and related languages.

Keywords: Hungarian National Corpus, spelling advisory portal, digitization of dictionary cards, Khanty and Mansi text corpora.

Prószéky Gábor – Váradi Tamás
HUN-REN Nyelvtudományi Kutatóközpont