A 75 éves Nyelvtudományi Kutatóközpont projektjei a magyar nyelv digitális támogatásáért és fenntarthatóságáért

Szövegerdőben bolyongva: lásd a fák helyett az erdőt!
Héja Enikő kutatócsoport-vezető, tudományos munkatárs (HUN-REN NYTK Nyelvtechnológiai és Alkalmazott Nyelvészeti Intézet Nyelvtechnológiai kutatócsoport)

Az előadásban bemutatjuk, hogy mi egy nyelvi korpusz, és hogy milyen tulajdonságokkal kell rendelkeznie ahhoz, hogy megfelelő kiindulási alap legyen empirikus nyelvészeti kutatások számára. Kitérünk arra, hogy miért fontos, hogy a korpusz nagyméretű legyen és sokféle szöveget tartalmazzon. Ismertetjük, hogyan segíthet egy jól összeállított korpusz a nyelvi mintázatok, gyakorisági eloszlások és jelentésváltozások feltárásában. Bemutatjuk a különböző típusú korpuszokat, az építésük során felmerülő kihívásokat, valamint az automatizált eszközök szerepét az adatok elemzésében. Az előadás célja, hogy rávilágítson, miként járulnak hozzá a nagyméretű korpuszok a nyelvi jelenségek mélyebb megértéséhez, és hogyan támogatják az empirikus leíró nyelvészetet.

Hány „Tisztelt Elnök Úr!”-at bír el egy korpusz?
Ligeti-Nagy Noémi tudományos munkatárs (HUN-REN NYTK Nyelvtechnológiai és Alkalmazott Nyelvészeti Intézet Nyelvtechnológiai kutatócsoport)

A Magyar Nemzeti Szövegtár 3.0 egy 10 milliárd tokent tartalmazó nemzeti referenciakorpusz lesz, amely jelentős méretbeli és lefedettségbeli bővülést jelent az 1 milliárd tokent tartalmazó MNSZ2-höz képest. A korpusz összeállítása számos kihívást tartogat. A szöveggyűjtés során az egyik kulcskérdés a doménspecifikus deduplikáció, különösen a sajtó stílusrétegen belül, ahol az ismétlődő tartalmak komoly szűrést igényelnek. Mivel kiemelt célunk, hogy a határon túli nyelvhasználat megfelelő súllyal legyen képviselve a korpuszban, olyan érdekes kérdésekben is döntést kell hoznunk, mint például, hogy egy Bécsben klinikát nyitó magyarországi orvos honlapja vajon határon túli nyelvhasználatot reprezentál-e.

Cédulák, adatbázisok és mesterséges intelligencia: a tudományos szakirodalom feltárása szakmai-könyvtári együttműködésben
Holl András informatikai főigazgató-helyettes (MTA Könyvtár és Információs Központ)

A tudományos szakirodalom vagy a levéltári forrásanyag feltárásának fontos lépése volt valaha a cédulázás. Egyéb források mellett a cédulaanyagra támaszkodva nagy lexikális munkák készültek, mint a *Magyar írók élete és munkái*, *A magyar irodalom története* vagy a *Magyarország régészeti topográfiája*. Az 1980–90-es években megjelentek a számítógépes adatbázisok a forrásanyag feltárásának eredményeként. A szakirodalom humán indexelése egyre nehezebben megoldható. Azonban a feldolgozást végző intelligencia lehet mesterséges is: a nagy mennyiségű digitális szöveg gépesített „kicédulázása” a könyvtárak és a szakmai közösségek együttműködésének terepe lehet.
A TMNP keretében, a HUN-REN NYTK és az MTA KIK nagy nyelvmodellek alkalmazásával dolgozza fel a REAL repozitórium modern magyar szöveganyagát.

A mesterséges intelligencia felhasználásának lehetőségei a helyesírási tanácsadásban

Ludányi Zsófia; *Váradi Tamás*; *Kocsis Ágnes*; *Madarász Gábor* (HUN-REN Nyelvtudományi Kutatóközpont)

Az MTA Nyelvtudományi Intézete 2013-ban nyitotta meg a Helyesiras.mta.hu online helyesírási tanácsadó portált, amely nyelvtechnológiai eszközök segítségével nyújt gyors, automatikus választ a felhasználók helyesírási kérdéseire. Számot vetve az indulás óta eltelt több mint tíz év tapasztalataival, időszerűvé vált a helyesírási portál korszerűsítése.
A mesterséges intelligencia alkalmazásával hatékonyabbá és egyszerűbbé tehető a portál használata. Terveink szerint az új felületen a felhasználók természetes nyelven fogalmazhatják majd meg helyesírási kérdéseiket, nyelvi problémáikat. Az interaktív technológia képes a kérdéses szóalakok nyelvi kontextusának pontosabb felismerésére, így a felhasználók az adott kérdéshez igazított, közérthető választ kaphatnak. Az új technológia a kutatóközpont nyelvi közönségszolgálatának tudásbázisára épít.

A magyar nyelv nagyszótára **archivális cédulagyűjteményének digitalizálása**
Simon László kutatócsoport-vezető, tudományos munkatárs (HUN-REN NYTK Lexikológiai Intézet Lexikai tudásreprezentáció kutatócsoport)

A túlnyomórészt A6-os méretű és kézzel megírt lapokból álló kollekció annak a „közösségi gyűjtés”-nek az eredményeként jött létre, amelyet a 19. század végén a *Magyar Nyelvőr* hasábjain kezdeményeztek. Az anyag nagyságát, a cédulák számát illetően az elmúlt 120 évben mindvégig csak becslések álltak rendelkezésünkre: a digitalizálási munkálatok következtében azonban ma már pontos információink vannak többek között arról is, hogyan oszlik meg az anyag az 1128 tárolódobozban.
Jelenleg a Magyar Nemzeti Levéltárral folytatott együttműködés keretében azon dolgozunk, hogy minden egyes cédulaképhez hozzárendeljük azt az információt, hogy az eredeti cédula melyik címszó illusztrálásához készült. Azt gondoljuk, hogy a munkálatok befejeztével az anyagot online is könnyen és jól kereshetővé tudjuk majd tenni.

Magyar terminológiastratégia
Lipp Veronika intézetigazgató (HUN-REN Nyelvtudományi Kutatóközpont Lexikológiai Intézet)

2023. december 1-jén indult a Magyar terminológiastratégia alprogram a Tudomány a Magyar Nyelvért Nemzeti Program keretén belül. A HUN-REN Nyelvtudományi Kutatóközpont által koordinált program célja a magyar terminológiastratégiai munkák belföldi és külhoni koordinációja. Ezenkívül kiemelt feladat a határon túli nyelvészeti kutatóállomásokat tömörítő Termini Magyar Nyelvi Kutatóhálózat működésének összehangolásával egy ingyenesen hozzáférhető, nyolc nyelven elérhető digitális oktatásterminológiai adatbázis létrehozása, egyben egy olyan közös magyar terminológiai adatbázisé, amely segíti többek között az oktatás szaknyelvének Kárpát-medencei harmonizációját. Az előadás az elmúlt egy évben elért eredményekről számol be.