Szerző: Horváth Dániel

Tízmilliárd szó a magyar nyelvért

A Magyar Tudományos Akadémia egyik új nemzeti programja a Tudomány a Magyar Nyelvért Nemzeti Program, amelynek célja a magyar nyelv kutatása a 21. században, amikor szinte minden tudományos tevékenységet áthat a digitalizáció és a mesterséges intelligencia. **Prószéky Gáborral,** a Nyelvtudományi Kutatóközpont főigazgatójával, a Tudomány a Magyar Nyelvért Nemzeti Program egyik vezetőjével beszélgettünk a program célkitűzéseiről és a többi között arról is, hogy miről beszél egy az *Egri csillagok*nál százezerszer hosszabb szövegkorpusz.

Mi a Tudomány a Magyar Nyelvért Nemzeti Program fő célkitűzése, miért volt szükség a létrejöttére?

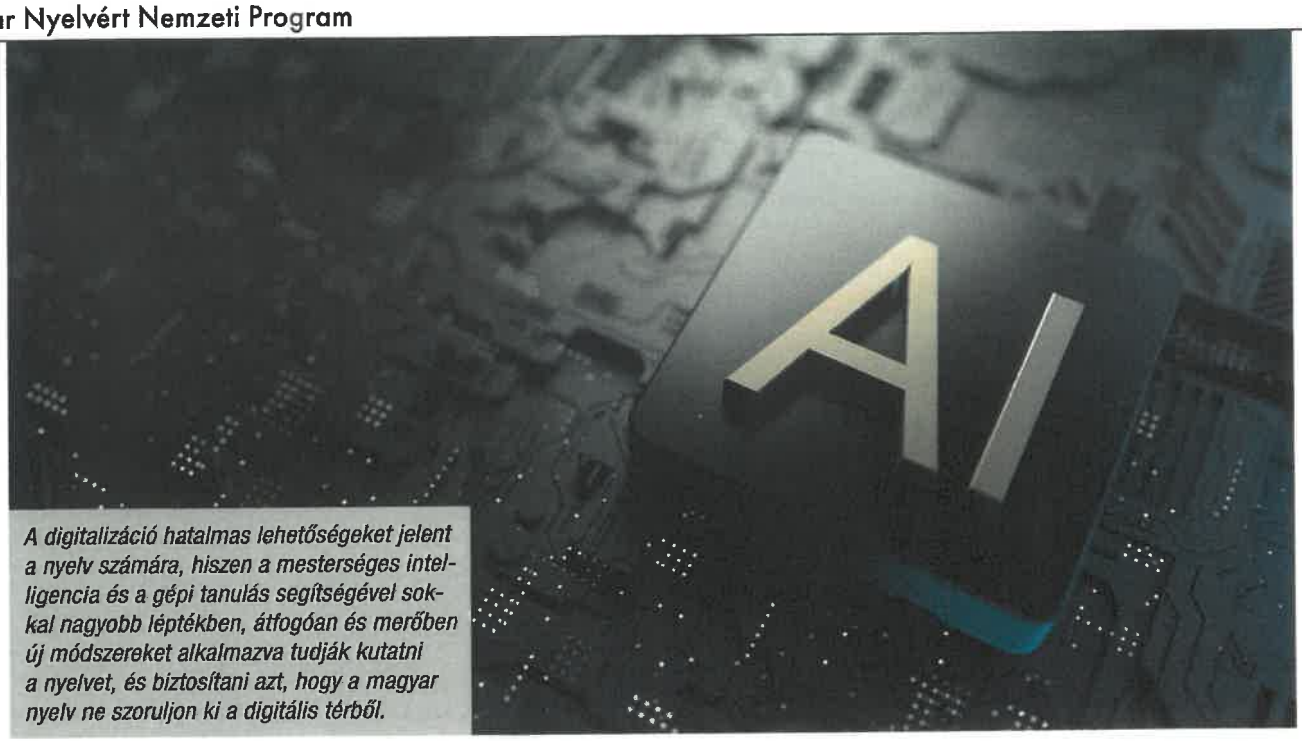
– Ez a program négy alprogramból áll, ebből a Nyelvtudományi Kutatóközpont két alprogramot koordinál.

Mindkét alprogram azt járja körül, hogy hogyan kezeljük nemzeti nyelvünket a digitalizáció világában.

Minthogy a Magyar Tudományos Akadémia alapításkor megfogalmazott küldetése szerint egyik fő feladata a magyar nyelv kutatása, e nemzeti program újraértelmezi és aktualizálja ezt a küldetést a 2020-as éveknek megfelelően. Minden nemzet elemi érdeke, hogy nyelve minél hatékonyabb formában legyen jelen a digitális térben. Meg kell jegyezni, hogy a magyar ebből a szempontból nem áll rosszul sok más nyelvhez képest, de ez nem jelenti azt, hogy ne lenne vele tennivaló.

Mi kell ahhoz, hogy egy nemzeti nyelv sikeres legyen a digitalizáció korszakában?

– Ez rendkívül sok tényezőtől függ, de fontos, hogy a mobil eszközökön elérhető digitális szolgáltatások az adott nemzeti nyelven is használhatóak legyenek, a nemzeti nyelvek fennmaradását erősen segítik például a nemzeti nyelvű wikipédiák is. A digitalizáció korántsem fenyegetést jelent a nyelvre, hanem hatalmas lehetőségeket, hiszen a mesterséges intelligencia és a gépi tanulás segítségével sokkal nagyobb léptékben, átfogóan és merőben új módszereket alkalmazva tudjuk kutatni a nyelvet, és biztosítani azt, hogy a magyar nyelv ne szoruljon ki a digitális térből. A mesterségesintelli-



A digitalizáció hatalmas lehetőségeket jelent a nyelv számára, hiszen a mesterséges intelligencia és a gépi tanulás segítségével sokkal nagyobb léptékben, átfogóan és merőben új módszereket alkalmazva tudják kutatni a nyelvet, és biztosítani azt, hogy a magyar nyelv ne szoruljon ki a digitális térből.

gencia-alapú nyelvészeti kutatáshoz – ezen belül például a magyar nyelv nemrég elkészült neurális modelljéhez – azonban jelentős mennyiségű szövegre van szükség, így a nemzeti program egyik célkitűzése e szövegállomány számottevő bővítése is.

Mélyreható elemzések

Honnan tudnak szövegeket szerezni a kutatásaikhoz?

– A legnagyobb nehézség, hogy bár a normatív, vagyis javított, lektorált, a helyesírási szabályoknak mindenütt megfelelő szövegek a legértékesebbek számunkra, ebből sokkal nehezebb nagy mennyiséghez jutni. Természetesen a „legolcsóbb" az lenne, ha az internetről gyűjtenénk szöveget, ami az internet nyelvét leíró modellek számára fontos is.

De a Nyelvtudományi Kutatóközpont és a Magyar Tudományos Akadémia számára kiemelkedő jelentőségű, hogy a gondozott, azaz a helyesírási szabályokat betartó magyar szövegekkel kapcsolatban gyűjtsünk ismereteket.

Hiszen ha ez megvan, azt már akár nyelvjárások, határon túli változatok, de az internet nyelvezetének a vizsgálatára is lehet használni, ha megfelelő módosításokat alkalmazunk rajta. Ám ehhez előtte meg kell alkotnunk a minél átfogóbb alapmodellt. Ezen dolgozunk már régóta, és ennek érdekében kezdtünk együttműködésbe az MTA Könyvtár és Információs Központjával, azaz a Magyar Tudományos Akadémia könyvtárával. Ennek az alprogramnak a címe: *A magyar nyelv digitális támogatása a magyar tudományosság szolgálatában.* A könyvtárban jelentős mennyiségű magyar szöveget őriznek, amelyek jól használhatók a modern nyelvtechnológiai kutatásokhoz, hiszen így rengeteg jó minőségű, autentikus szöveghez jutunk, amelyekből sokat tanulhatnak a modelljeink.

Miért fontos ez a kooperáció a könyvtár számára?

– Az együttműködés a könyvtárnak is rendkívül hasznos, hiszen egyik fontos feladatuk a REAL repozitórium, vagyis a könyvtári tételek teljes szövegeit tartalmazó szövegbázis kialakítása és gondozása. Most ennek az anyagát a nyelvtechnológia alkalmazásával tesszük a jelenleginél hatékonyabb módon kutathatóvá, és a tudományos publikációk óriási tömegének tartalmát tesszük könnyen kereshetővé a szövegekből automatikusan kinyert, eddig csak ember által megfogalmazott osztályozási ismérvek segítségével. Ezáltal jobban lehet benne később keresni, így elérhetőbbé válnak a benne esetleg elrejtett tartalmak. Ha használhatok egy ma divatos kifejezést, akkor ez egy „win-win" helyzet, amelyben mindkét fél nyer: a könyvtári szolgáltatások magasabb szintűek lesznek, miközben a szövegek elolvasása után a magyar nyelv modelljei is javulnak. A mai okos rendszerek ugyanis már korántsem csupán arra képesek, hogy egy szöveget karakterek sorozataként kezeljenek, és ennek alapján esetleg keresni tudjanak bennük. Ha nagyobb mennyiségű autentikus szövegből van lehetőségük tanulni, akkor sokkal mélyrehatóbb elemzéseket képesek elvégezni a szövegen.

Mivel foglalkozik majd a másik alprogram?

– A másik alprogram, amelyben részt veszünk, teljes egészében a Nyelvtudományi Kutatóközpont saját kutatási témája, címe: *A magyar nyelv digitális fenntarthatóságának támogatása.* A munkálatok több olyan területet fednek le, amelyeket az MTA kezdeményezett korábban, és most e programokat emeljük magasabb szintre. Mi kezeljük például az MNSZ-t, a Magyar Nemzeti Szövegtárat, amelybe folyamatosan és kiegyensúlyozottan gyűjtjük a szövegeket (amelyek a legkülönbözőbb forrásokból származnak, és eltérő típusúak). Ezen a hatalmas szövegbázison a háttérben futó programjaink segítségével a nyelvészek a korábbiaknál alaposabban adatolt kutatásokat tudnak folytatni, és így gazdagítják a magyar nyelvről meglévő ismereteinket, amit a köz javára bocsátunk. A szövegtár kezdetei immár negyed évszázadra nyúlnak vissza. Kezdetben 187 millió szóból állt ez a nyelvi korpusz, ami akkoriban

óriásinak tűnt, de a mai MNSZ2-ben már egymilliárd szónál tartunk. Ezt a méretet tervezzük tízszeresére emelni ennek az alprogramnak az egyik céljaként. Vagyis a fő cél az, hogy a mostaninál lényegesen nagyobb, rendszerezett, és a kutatók számára az eddigieknél is jobban elérhetővé tett nyelvi korpuszt hozzunk létre.

A beszélő szövegtest

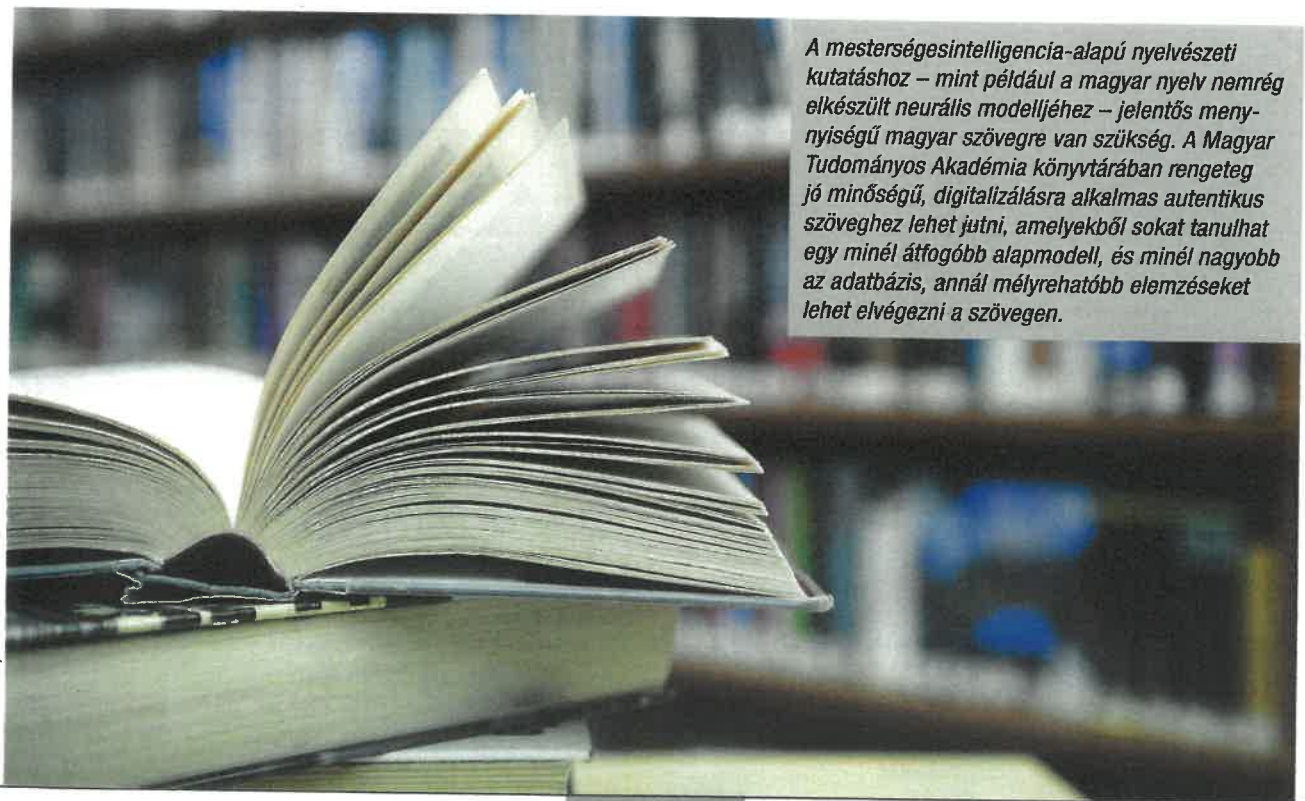
Mi jellemzi a jól kutatható szövegkorpuszt?

– A korpusz az a szövegtest, amelyet egy-egy nyelv korábban teljes egészében megalkotott szövegeinek egymás mellé helyezésével és előkészítésével hozunk létre. A korpuszt olyan szövegek alkotják, amelyeket egy az adott nyelven tudó személy autentikus formában írt le, idetartoznak a szépirodalmi szövegek, de az interneten megtalálható szövegek, az újságok szövegei, illetve a könyvek szövegei is. Fontos, hogy a korpusz teljes egészében, elejétől a végéig tartalmazza az eredeti szövegeket, mert azokból nem csupán a szavak a lényegesek, hanem a teljes szöveg felépítése, a különböző szövegtípusok eltérő jellegzetességei is. A szövegkorpuszok a hagyományos értelemben arra valók, hogy bennük a kutatók mintázatokat fedezhessenek fel. De a modern matematikai módszerek segítségével ma már maga a korpusz is képes „beszélni", vagyis, ha elegendően nagy a szövegtest, akkor olyan törvényszerűségek, egyezések, hasonlóságok bukkanhatnak fel a korpuszból, amikre a nyelvész talán nem is gondolt. A korábbi, ugyancsak jelentős, de a maiaknál sokszor kisebb nyelvi adatbázisok ilyenfajta kutatásra alkalmatlanok voltak. A világ legfontosabb nyelveit kutató nyelvészeti vizsgálatok alapvetően ebbe az irányba mennek, és nagyon örömteli, hogy most nekünk is lehetőségünk van ide eljutni.

A laikusok talán a helyesírási problémák esetén találkoznak leggyakrabban a nyelvészettel. A helyesírásnak is szerepe lesz a programban?

A helyesírás egy nyelvvel kapcsolatos társadalmi megállapodáson nyugszik, de nem a nyelv alapvető tulajdonsága.

Ettől függetlenül az élet számos területén játszik fontos szerepet, hiszen a kommunikáció egységességét szolgálja, ehhez pedig kell a norma, ami szerint leírjuk a mondandónkat. Bár a helyesírást sokszor összekeverik a nyelvészettel, a kettő nem azonos, jóllehet kétségtelen, hogy a helyesírási szabályokat is nyelvészek alkotják, nyelvészeti elvek alapján. Minthogy a mesterséges intelligenciát a helyesírás pontosabb alkalmazására is használhatjuk, a nemzeti program részeként meg fogjuk újítani a *helyesiras.mta.hu* címen immár sok éve elérhető helyesírási tanácsadó szolgáltatásunkat. Ez az alapításakor igen korszerű megoldásokat tartalmazott, de ahogy minden informatikai alapú tudományág, úgy a számítógépes nyelvészet is szélsebesen fejlődik, és ennek megfelelően fokozatosan elavulnak a korábban létrehozott eszközök. E fejlesztés eredményeként a mostaninál lényegesen rugalmasabb, intelligensebb rendszer fog létrejönni, amely sokkal gyorsabban megérti majd, hogy pontosan mi a felhasználó problémája, és erre releváns választ fog neki adni. Egyébként a Nyelvtudományi Kutatóközpontnak vannak más hosszú távú projektjei is, amelyek bizonyos elemeikben akár a 19. századig is visszanyúlhatnak. Ilyen például a magyar nyelv nagyszótára, amelyet hosszú kihagyás után húsz évvel ezelőtt elkezdtünk felépíteni, és e munka jelenleg is folyik. A nagyszótár készítése során az elmúlt évtizedekben kollégáink az egyes feldolgozandó szavakat tartalmazó cédulák millióit hozták létre. Ezek fizikailag is létező cédulák, kézzel írt kártyák, amelyeket intézményünkben őrzünk. Ezeket a most induló projekt keretében digitalizálni fogjuk, hogy ez a cédulatömeg a maga teljességében elérhetővé váljon a jövő kutatói számára. Ha ez digitálisan is hozzáférhető lesz, akkor nem feltétlenül kell hatalmas



A mesterségesintelligencia-alapú nyelvészeti kutatáshoz – mint például a magyar nyelv nemrég elkészült neurális modelljéhez – jelentős mennyiségű magyar szövegre van szükség. A Magyar Tudományos Akadémia könyvtárában rengeteg jó minőségű, digitalizálásra alkalmas autentikus szöveghez lehet jutni, amelyekből sokat tanulhat egy minél átfogóbb alapmodell, és minél nagyobb az adatbázis, annál mélyrehatóbb elemzéseket lehet elvégezni a szövegen.

termeket fenntartani a gyűjtemény számára, vagyis költséghatékonyabb megoldás is – emellett pedig kulturálisörökség-mentésként is felfogható.

Csak a magyar nyelvű korpusz fejlesztése a nemzeti program célja?

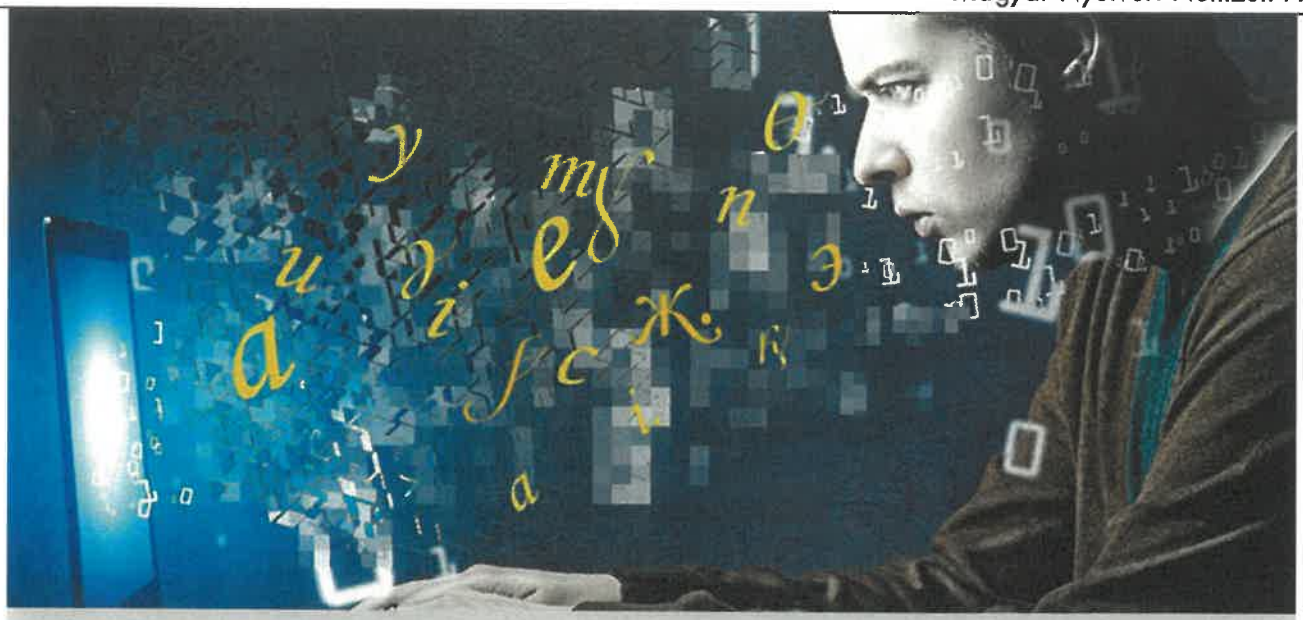
– A magyar nyelv legközelebbi rokonai a hanti és a manysi nyelvek, amelyek ma élő beszélői Oroszországban élnek. A magyar nyelvészeti kutatások számára korábban sem voltak könnyedén elérhetők a hanti és a manysi források, de az ukrajnai háború kitörése óta ezek a személyes részvételt igénylő vizsgálatok gyakorlatilag ellehetetlenültek. Viszont adnak ki folyóiratokat ezeken a nyelveken, így ezekből igyekszünk a magyar nyelvi szövegtárhoz hasonló – bár annál nyilván jóval szerényebb – korpuszt létrehozni hanti és manysi nyelvekre is. Ilyet még soha nem alkotott senki. E nyelveknél különösen érdekes kérdés, hogy írott formájuk mennyiben különbözik a beszélt változattól. Noha létezik hanti és manysi írásbeliség, de ezek fenntartása számos mesterséges beavatkozást kíván, hiszen a folyóiratok szövegében található egyes, mondjuk úgy, hogy szaknyelvi elemek nagy valószínűséggel nincsenek használatban a köznapi beszélt nyelvben. Hasonlóan a magyar nyelvi korpuszra fejlesztett rendszerhez, hantira és manysira is elkészítjük majd a korábban említett nyelvmodelleket. Az sem kizárható, hogy a magyar nyelvre fejlesztett modell módosítva – mintha csak egy nyelvváltozat lenne – alkalmazható lehet a struktúrájában hasonló hanti és manysi korpuszokra is, és így remélhetően ezek segítségével jelentősen gyarapodni fog az e nyelvekkel kapcsolatos tudásunk. Úgy gondoljuk, hogy valahol a mi felelősségünk, hogy a magyar nyelv kutatására kapott támogatásból segítsük e kevesebb támogatásban részesülő rokon nyelvek vizsgálatát is. E nyelvek ugyanis nemcsak a hagyomány vagy valamilyen önkényes döntés következtében minősülnek a magyar nyelv rokonainak, hanem tudományosan bizonyított tény, hogy közös az eredetük.

Óriási korpuszok

Miben jelent pluszt a modern gépi feldolgozás a hagyományos nyelvészeti kutatómunkához képest?

– Ha a korpuszt alkotó szövegeket egyéb kiegészítő információkkal látjuk el, szakszóval: annotáljuk, azaz például megjelöljük, hogy mely tájegységről származik, mi a témája, ki a szerzője, akkor az algoritmusok segítségével a karaktersorozat mögött rejlő mélyebb rétegekre is rá tudunk majd kérdezni. Ha ezt emberi erővel akarnánk megvalósítani, az felmérhetetlenül sok munkát igényelne. A mesterséges intelligencián alapuló modellek már egészen kifinomult nyelvtani elemzésekre is képesek, meg tudják találni a mondatrészeket, a szófajokat és sok más jellegzetességet is azonosíthatnak. Manapság már csak meg kell tanítani a gépet arra, hogy mit kell keresnie, majd az elvégzett munkát vissza kell ellenőriznünk. Az ellenőrzött szövegből újra tanul, így a következő

A mesterséges intelligenciát a helyesírás pontosabb alkalmazására is használhatjuk. A nemzeti program részeként megújul a helyesiras.mta.hu; *a mostaninál lényegesen rugalmasabb, intelligensebb rendszer jön majd létre, amely sokkal gyorsabban megérti, hogy pontosan mi a felhasználó problémája, és erre releváns választ fog neki adni.*



A szövegértés komplex művelet, amelyhez nemcsak nyelvtudás, hanem a világról szóló hétköznapi tudás is szükséges. A mesterségesintelligencia-alapú programok valójában buták, hiszen nem a mi logikánk szerint építik fel magukban a szövegből kibontható világot. Viszont ezeknek a programoknak a reakciói sokszor az általuk „soha nem látott" világ általunk is ismert, de az egyes szövegekben explicit módon mégsem megjelenő felépítésére utalnak.

körben már jobb eredményt ad. Ezt is kijavítjuk, és így fokozatosan egyre hatékonyabb elemzést lehet elérni. Ez a folyamat nyilván összehasonlíthatatlanul gyorsabb, mintha az embernek kellene elvégeznie kézzel az elemzést. Így nagyságrendekkel több szöveget lehet emberi időléptékben annotálni, ami alapján sokkal átfogóbb és megbízhatóbb eredményre vezető vizsgálatokat lehet végezni a korpuszon. Egy több milliárd szavas korpusz mérete minden képzeletet felülmúl. Csak összehasonlításul: az *Egri csillagok* talán 135 ezer szóból áll.

A digitális nyelvészet tehát alapvetően a kutatók kapacitását növelte meg, vagy minőségi változást is hozott?

Megkockáztatom, hogy a nyelvészet digitalizációja gyökeres változásokat eredményezett, de nem szeretnék bombasztikus szavakat használni.

Önmagában a digitalizáció először csak megnövelte a hatékonyságot, és lehetővé tette, hogy kiterjedtebb korpuszalapú kutatást végezhessünk. Korábban az volt az általános, hogy egy tudományos munkához a kutató legfeljebb néhány száz nyelvi szerkezetet tudott összegyűjteni, és azon végzett elemzéseket. A mai korpuszok korában egy szempillantás alatt talál a gép nagyságrendekkel több példát, így teljesen másféle vizsgálatokat tudunk végezni – már csak azért is, mert egyszerűen marad rá időnk. De a legutóbbi időkben tényleges minőségi ugrás történt a nyelvészetben. A mai óriási méretű korpuszok nem csak emberi kutatásra szolgálnak, színre lépett a gépi tanulás, amellyel a csak néhány éve létrehozott mesterséges neuronhálókon alapuló modellek olyan jellegzetességeket tárnak fel a nyelvből, amelyek létezését nem is sejtettük. A mesterséges intelligencia működtetése során az derült ki, hogy leírt szövegeink, azaz a karaktersorozatok olyan információt is tartalmaznak, amely sokkal gazdagabb a puszta alaktani, mondattani vagy más nyelvészeti kategóriáknál.

Ez a „többletinformáció" hogy érhető tetten a nyelvhasználatban?

– Jó példa erre, hogy bár a magyar nyelvben nincs nemük a személyes névmásoknak, ezek a modellek szinte tévedhetetlenül megállapítják egy elegendően hosszú magyar szövegről, hogy az nőről vagy férfiról szól – annak ellenére, hogy ez az általunk ismert nyelvi elemekből nem derül ki.

Vagyis jóval nagyobb tudás rejlik ezen egyszerű betűsorozatokban a szintaxisnál és a nyelvészek által eddig tanulmányozott jellemzőknél.

Rekonstruálható belőlük az a világismeret, amely a beszélő fejében megvan, de az korántsem volt egyértelmű eddig, hogy mindez a tudás valahogy belekerül a szövegbe is. E rendszerek – vagy akár a fordítóprogramok – azért „okosak", mert már nem a betűsorozatot értékelik, hanem felismerik a szöveg rejtett mintázatait is.

A mesterséges intelligencia érti a szöveget a szó emberi értelmében?

– A szövegértés komplex művelet, amelyhez nemcsak nyelvtudás, hanem a világról szóló hétköznapi tudás is szükséges. E tekintetben ezek a rendszerek valójában buták, hiszen nem a logika, legalábbis nem a mi logikánk szerint építik fel magukban a szövegből kibontható világot. Viszont ezeknek a programoknak a reakciói sokszor az általuk „soha nem látott" világ általunk is ismert, de az egyes szövegekben explicit módon mégsem megjelenő felépítésére utalnak. Sőt az ezt működtető neurálisháló-alapú módszerek úgy képesek kommunikálni, mintha valóban volnának arról tapasztalataik, amiről „beszélnek". Ám ezeknek a programoknak nincs, nem is lehet ilyen tapasztalatuk, tehát amit csinálnak, az az emberek által létrehozott szövegekben megbúvó összefüggések komplex összemásolása, de sok esetben olyan módon, mintha tényleg tudnák is, miről van szó. ■