NYELVÉSZETI SZÖVEGKERESŐK, NEMZETI KORPUSZPORTÁL

Sass Bálint

PhD, MTA Nyelvtudományi Intézet Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály
sass.balint@nytud.mta.hu

1. Bevezetés, fogalmak, elvek

Mi az, hogy *nyelvészeti szövegkereső?* Olyan számítógépes eszköz, gyakran *online* felület, melynek segítségével szövegekben, általában nagy méretű szöveggyűjteményekben kereshetünk adott tulajdonságú szavakat, szókapcsolatokat, és a keresési feltételeket különféle nyelvészeti szempontok, a szavak nyelvészeti tulajdonságai alapján adhatjuk meg. Az ilyen eszközöket a (számítógépes) nyelvészetben *korpuszlekérdezőnek* szokás hívni, azért vezetem be a fogalomra szinonímaként a nyelvészeti szövegkereső megjelölést, mert ez talán jobban érthetővé teszi, hogy miről is van szó.

Mi a korpusz? Sok szöveg. Mi a lekérdezés? Valamiket keresünk ebben a szövegben. A korpusz fogalmát leszűkítjük annyiban, hogy csak elektronikusan tárolt szövegekről beszélünk, és ezen belül is csak a „géppel olvasható", karakteres formában rendelkezésre álló szövegekről. A képként tárolt szövegek tehát nem megfelelőek, szükséges, hogy a szöveg a számítógép számára azonosítható karakterek sorozataként legyen reprezentálva.

Mekkorák valójában ezek a nagy méretű szöveggyűjtemények? A szemléletesség kedvéért képzeljük el a korpuszokat könyvek formájában egy könyvespolcon. Egy 5 centi szélességű könyv nagyjából százezer szót tartalmaz. Eszerint egymillió szó 50 centis, százmillió szó 50 méteres, és a ma általánosnak tekinthető milliárd szavas korpuszok 500 méteres könyvsorként képzelhetők el.

Sok szövegben akarunk tehát keresni. Gondolhatnánk, hogy ez meg van oldva, rendelkezésünkre állnak a különféle internetes keresők, melyek éppen ezt csinálják, valóban nagyon sok szövegben képesek keresni, és a releváns információt visszaadni. Lassan egyetemista lesz az a korosztály, amelyiknek a születésekor már létezett a hatékony internetes keresés.

Adam Kilgarriff 2007-es *Googleology is Bad Science* – magyarul nagyjából *A guglizás mint nem megfelelő tudományos módszer* – című cikkében (Kilgarriff, 2007) összeveti az internetes keresőket a korpuszlekérdezőkkel, és bemutatja, hogy az előbbiek miért nem alkalmasak arra, hogy nyelvészeti szövegkeresőkként használjuk őket. Négy szempontot sorol fel, érdemes itt áttekinteni ezeket. Az internetes keresőkben (1) nincs nyelvészeti adatolás, mint például az egyes szavakhoz a szótő vagy a szófaj; (2) a keresőkifejezést csak nagyon behatárolt módon lehet megadni; (3) korlátozva van a naponta futtatható lekérdezések száma (ez akkor probléma, ha automatikusan szeretnénk futtatni sok, akár több ezer lekérdezést) és nem kapjuk meg az összes találatot; valamint (4) egy találat egy dokumentumot (internetes oldalt) jelent, nem pedig egy szóelőfordulást.

Az internetes keresők tehát nem nyelvészeti szövegkeresők, a nyelvészeti adatolás hiánya, illetve az egyszerű lekérdezési formátum miatt nem tudunk nyelvészetileg releváns lekérdezéseket megfogalmazni. Az internetes keresők ún. információ-visszakereső (*information retrieval* – IR) rendszerek, feladatuk az, hogy azt a dokumentumot adják eredményül, mely a lekérdezésnek megfelelő releváns információt tartalmazza. A nyelvészeti szövegkeresők feladata ezzel szemben az, hogy adott, precízen körülírt nyelvi jelenség *összes* előfordulását szolgáltassák, így nemcsak példákat kapunk a jelenségre, hanem lehetővé válik a jelenség statisztikai vizsgálata is.

Nézzük meg, hogy mit kell tudniuk a nyelvészeti szövegkeresőknek a fenti négy szempont tekintetében.

A szövegek nyelvi adatolást – ún. *annotációt* – kell, hogy tartalmazzanak. Az annotáció azt jelenti, hogy a szöveg egyes egységeihez különféle adatok vannak rendelve. Például a dokumentumokhoz a szerző vagy a szavakhoz a szófaj, de a bekezdésekhez, mondatokhoz is társíthatók adatok, mint például az adott egység nyelve vagy mérete. Az annotációk általában valamilyen egységes kódrendszer szerint szerepelnek a korpuszokban.

Példa: A *körülültük* szóalak morfológiai annotációja, azaz ami a szó alaktanát, a benne lévő elemeket írja le a következő lehet: `IK.IGE.TMt1`. Ebből kiderül, hogy ez a szóalak tehát egy igekötős (`IK`) ige (`IGE`), mely határozott ragozású (`T`) – vö: *körülültük az asztalt* és nem *egy asztalt* –, múlt idejű (`M`), és többes szám első személyű (`t1`).

Az annotációk kézi, gépi (automatikus) vagy félautomatikus (gépi annotálás + kézi ellenőrzés) úton kerülnek bele a korpuszokba. Számos nyelvre számos automatikus szótövező, morfológiai elemző, szintaktikai elemző stb. eszköz létezik. Nagyobb szövegeknél és megbízható gépi eszközök esetén kap teret az automatikus gépi annotálás. A korpusz fogalmát a fentiek alapján tovább szűkíthetjük az annotált szövegekre.

A nyelvészeti szövegkeresők fontos tulajdonsága, hogy az alapegység nem a dokumentum, hanem leggyakrabban a szó. Esetleg lehet más, dokumentumnál kisebb egység is: mondat, tagmondat vagy akár a hang, de mindenképpen valamilyen nyelvészeti szempontból releváns egység. A filológiával, irodalomtudománnyal szemben, mely talán nagyobb jelentőséget tulajdonít a dokumentumokhoz rendelt, dokumentumszintű adatoknak (szerző, forrás, megjelenés ideje stb.), a nyelvészeti szövegkeresőkben a szó a központi elem. Egyrészt általában a szavak kapják a legtöbb fajta és legrészletesebb annotációt, másrészt a találatok sem dokumentumok, hanem szavak. Másképp fogalmazva: egy nyelvészeti szövegkereső esetében alapvető követelmény, hogy a keresett szó minden egyes előfordulását külön találatként jelenítse meg.

Az irodalomtudományban ma elterjedt Franco Moretti-féle *distant reading* (távoli olvasás) paradigma (Moretti, 2013) több mű, sok szöveg („adat") aggregált (statisztikai) vizsgálatát javasolja, szembeállítva az egyes művek mélyreható vizsgálatát, aprólékos tanulmányozását jelentő hagyományos *close reading* (közeli vagy szoros olvasás) iránnyal. Egy irodalmi mű tanulmányozásához a mű elolvasása mindenképpen szükségesnek tűnik, Moretti mégis lényegében azt javasolja, hogy ne olvassuk el a műveket. Úgy tűnhet, hogy a fenti szóközpontú nyelvészi megközelítés a hagyományos közeli olvasás irányt képviseli.

Úgy véljük, hogy ez nincs így: a nyelvi információk mindkét megközelítésben segíthetik a kutatást, a távoli olvasás jellegű statisztikai vizsgálatokhoz éppen az annotált, elemzett korpuszok szolgáltathatják a gondosan előkészített, tiszta nyelvi adatot.

Úgy is mondhatjuk, hogy a nyelvészeti szövegkeresők esetében szeretnénk pontosan megadni, hogy *hol* keresünk és hogy *mit* keresünk. Egyrészt fontos, hogy mi a szöveganyag, azaz össze kell állítani az aktuális kívánalmak, kutatási kérdések szerinti korpuszt, legyen az egy sajtókorpusz, egy adott regény, Petőfi Sándor összes műve, Kovács Pisti Facebook-bejegyzései vagy a magyar nyelvet egészében jól reprezentáló nagyméretű korpusz. Másrészt, ahogy fentebb is írtuk, fontos, hogy meghatározzuk, meghatározhassuk, hogy pontosan mit keresünk. Nyelvészetileg releváns kérdéseket szeretnénk feltenni. Nyelvi tudású keresőt szeretnénk, ami adatot szolgáltat a magyar nyelv, a magyar nyelvű szövegek vizsgálatához.

A korpuszlekérdező tehát olyan számítógépes rendszer, mely meghatározott, alkalmasan annotált szöveganyagon, nyelvészetileg releváns kérdésekre tud válaszolni. A korpuszban rejlő nyelvi tudást a korpusz annotációja tartalmazza, ez teszi lehetővé, hogy nyelvészeti szempontok szerint pontosan megadhassuk, hogy mit keresünk.

Miért szükséges a korpuszlekérdező a nyelvészeti munkákhoz? Valós nyelvi megnyilatkozások gyűjteményeként a korpusz az, ami a hiteles nyelvi adatot szolgáltatja a kutatási kérdések megválaszolásához, a nyelvészeti hipotézisek alátámasztásához, illetve cáfolatához. A megfelelő korpusz objektívebb tud lenni, mint a nyelvész intuíciója vagy a korábbi évtizedekben használt célzott „cédulázós" kézi adatgyűjtés.

A nyelvészeti szövegkeresők legfontosabb sajátossága talán abban ragadható meg, hogy nemcsak adott szavakra, hanem nyelvészeti szempontok szerint megadott *szóosztályokra* is kereshetünk a segítségükkel. Adott lekérdezésre kapott válaszban általában nem egy konkrét szó előfordulásai, hanem a megadott feltételeknek megfelelő szóosztály tagjainak előfordulásai szerepelnek.

Példák: Ha a morfológiai annotációnál bemutatott példa szerint igekötős, határozott ragozású, múlt idejű, többes szám első személyű igéket keresünk, akkor a *körülültük, felszedegettük, elsimítottuk, végigcsináltuk…* szóosztály tagjait fogjuk megkapni eredményként. Hasonlóan, ha *f*-fel kezdődő *-ban/-ben* ragos többes számú főneveket keresünk, akkor a *forrásokban, fellegekben, falvakban, fejekben…* szavakat találjuk. A fenti morfológiai (alaktani) példák után vegyünk egy fonológiai (hangtani) példát is harmadikként: ha a lekérdezés olyan három hangból álló szótövekre irányul, melyek első hangja ún. affrikáta (*c*, *dz*, *cs* vagy *dzs*), a második hang tetszőleges magánhangzó, a harmadik pedig ún. approximáns (*l* vagy *j*), akkor *cél, csal, csaj, csel, dzsal…* lesz az eredmény.

A nyelvészeti szövegkeresőkben a legtöbb esetben nemcsak egyes szavakra, hanem szavak sorozataira, szókapcsolatokra is lehet keresni, következésképpen szóosztályok sorozataira, szóosztályok kapcsolataira is, amint erre a 2.5. részben példát is fogunk látni. Az összes találatnak köszönhetően megtudjuk az egyes szavak, szókapcsolatok gyakorisági adatait és viszonyait, valamint legtöbbször lehetőség van az egyes találati szavak környezetének vizsgálatára is.

E tanulmány két nagyobb részből áll. A következő részben bemutatjuk az elmúlt több mint tíz évben, a Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztályán készült korpuszokat és nyelvészeti szövegkeresőket. Utána pedig arról fogunk gondolkozni, hogy hogyan lehetne a jövőben ezt a sokféle korpuszt és sokféle lekérdezőt valahogyan egységes keretben kezelni, egységes keretben prezentálni, sőt reklámozni a potenciális felhasználók felé. Erről fog szólni a Nemzeti Korpuszportál ötlete.

2. Korpuszlekérdezők evolúciója

Ebben a részben öt nyelvészeti szövegkeresőt fogunk bemutatni. A cél az, hogy illusztráljuk az effajta technológiai eszközök fejlődését az évek során megjelenő újabb funkciók révén. A bemutatás nem teljes körű, csak néhány kiválasztott érdekes, hasznos funkcióra terjed ki.

2.1. MNSZ1 • Az első jelentős számítógépes nyelvészeti szövegkereső rendszer, amely az MTA Nyelvtudományi Intézetében készült, a Magyar Nemzeti Szövegtár első változata (MNSZ1) volt (Váradi, 2002). Jelen formájában 2005 óta érhető el, közel kétszázmillió szónyi szöveganyagot foglal magában. A magyar nyelv reprezentatív korpusza kíván lenni, ezért különféle stílusrétegekből (szépirodalom, hivatalos, tudományos, sajtó, internetes fórumok) ölel fel a hazain kívül határon túli anyagot is.

Az *1. ábrán* láthatók a *fut* szótőből és az azt maximum öt szó kihagyással követő *után* névutóból álló szókapcsolat előfordulásai ún. *konkordancia* formátumban. Ez a klasszikus megjelenítési forma azt jelenti, hogy a találati szavakat kiemelve egymás alá rendezzük, és a környezetükkel együtt mutatjuk be.

Kilgarriff első és második szempontjának megfelelően elkülönítjük a korpuszlekérdezők annotációs funkcióit és lekérdezőfunkcióit. A bemutatott korpuszokat e két szempontból vizsgáljuk: egyrészt a korpuszban meglévő annotáció (nyelvi tudás) jellegzetességeit tekintjük, másrészt pedig a lekérdezőfelületen elérhető eszközöket.

1.	szórt a gyepre. Szinte	**futottam** *V.Me1*	a biztos pont **után** *NU*. Az ablak alatt megálltam
2.	tízfillérest, Bambó meg röhögve	**futott** *V.Me3*	a guruló pénz **után** *NU*, míg csak ormótlan cipőjével
3.	formának tekinteni. Nem kell	**futnia** *V.INRe3*	a maszek munkák **után** *NU*, minden figyelmét a zenekari
4.	árak újra estek, mindenki	**futott** *V.Me3*	a pénze **után** *NU*, ami már rég nem
5.	Közben a Játékvezető Testület is	**fut** *V.e3*	a pénze **után** *NU*. Bár a versenykiírás szerint
6.	az, amikor egy szakember	**fut** *V.e3*	a pénze **után** *NU*. Hiszen azt talán maga
7.	, hogy azok, akik	**futnak** *V.t3*	a pénzük **után** *NU* ma inkább verőlegényekhez fordulnak,
8.	nem tettek mást, mint	**futottak** *V.Mt3*	a pénzük **után** *NU*. Az IT-Com vezette konzorcium
9.	legyen. Nehogy az értelmiség	**fusson** *V.Pe3*	a politika **után** *NU*. Mert igen fontos a
10.	őket Segesvárra. Éva hosszan	**futott** *V.Me3*	a vonat **után** *NU*. Egyik kezével integetett,
11.	Mellétérdepelve gyönyörködtem benne, aztán	**futottam** *V.TMe1*	apám **után** *NU*. (…) A
12.	a bámuló társnőjének, s	**fut** *V.e3*	át a doktor **után** *NU*. Tán a gyerekét gyógykezelte
13.	alapján a Győrnek kellett volna	**futnia** *V.INRe3*	az eredmény **után** *NU*. Győri ETO-Dunaferr 2-1 (
14.	ellenfél vezetett, mi csak	**futottunk** *V.Mt1*	az eredmény **után** *NU*. Néhányszor egyenlítettünk, de
15.	ezek után a Majoroséknak kellett	**futniuk** *V.INRt3*	az eredmény **után** *NU*. Némi mezőnyfölényre szert tettek
16.	tenni valamit, ám csak	**futottak** *V.Mt3*	az események **után** *NU*. Ezért célszerű végre átfogóan
17.	számlát kell ellenőriztetnem, s	**futottam** *V.Me1*	Judit **után** *NU*. Éppen letette a kagylót
18.	felfüggesztettük a patikazárlatot, mégis	**futnunk** *V.INRt1*	kell a pénzünk **után** *NU*. Tárgyalunk, de minden
19.	jövőre 2, 1 milliárddal	**fut** *V.e3*	ki. A zárás **után** *NU* fejeződik csak be az Uránia
20.	párizsi, brüsszeli életmódra nem	**futja** *V.Te3*	, másfajta vállalkozások **után** *NU* nézett. A szürkegazdaság fogalmát

1. ábra • A *fut… után* lekérdezés konkordanciája az MNSZ1-ből. A félkövér találati szavak mellett dőlt betűvel a morfológiai annotáció szerepel. Utóbbinak köszönhetően a *fut* összes ragozott alakját, (a *fut* alakjaiból álló szóosztályt) kapjuk meg. Látjuk, hogy általában a pénzünk után futunk, de azért néha a vonat, sőt Judit után is.

Már ezen a korai felületen megtaláljuk a legalapvetőbb, leggyakrabban meglévő annotációkat és funkciókat. Az annotáció tartalmazza a *szótövet* és a *morfológiai elemzést*, a lekérdező *konkordancia* formában jeleníti meg a találatokat, és képes *szókapcsolatokra* is keresni. A konkordancia megfelel annak a korábban megfogalmazott követelménynek, hogy a keresett szó, jelenség minden előfordulását külön találatként jelenítse meg.

2.2. Mazsola • A jelen formájában 2009 óta hozzáférhető Mazsola korpuszlekérdező (Sass, 2009) ugyanazon a szöveganyagon működik, mint az MNSZ1, de egy más szempontú keresési lehetőséget kínál kiegészítő annotációja révén.

A korpuszban itt meg vannak jelölve a tagmondatok, azonosítva van a tagmondat igéje (esetleges elváló igekötőjével együtt), és azonosítva vannak az ige mellett megjelenő névszói bővítmények (alany, tárgy, ragos és névutós névszók). Ez az annotáció teszi lehetővé, hogy a szórend fölött általánosítsunk, azaz igéket és bővítményeket, igékből és bővítményekből álló szerkezeteket a konkrét, az adott szöveghelyen épp megjelenő szórendjüktől függetlenül vizsgáljunk. Mondhatjuk, hogy ennek a lekérdezőnek az alapegysége nem a szó, hanem a tagmondat.

Példák: Olyan típusú lekérdezéseket fogalmazhatunk meg itt, hogy mik a *hagy* jellegzetes tárgyai, vagy, hogy mi a jellegzetes ige, ha a két bővítmény *hideg* és *hátán*. Az első esetben többek között a *kívánnivalót, nyomot, kétséget, üzenetet* szavakat kapjuk. Látjuk, hogy ezek nem egyszerű tárgyak, hanem a *hagy*-gyal speciális jelentésű szókapcsolatot (összetett igét) alkotó szavak. A Mazsola jellemző módon felfedi az efféle szerkezeteket. A második esetben a *végigfut, futkos, futkározik* igéket kapjuk *(2. ábra)*. A *Magyar szókincstárban* (Kiss, 1998) a szóban forgó szólásnak csak az első két változata szerepel. A Mazsola alkalmas kötöttnek vélt szerkezetek, szólások változatainak vizsgálatára is, a nagy korpusznak köszönhetően teljesebb képet kaphatunk a vizsgált jelenségről.

A Mazsola annotációs szintű újdonsága a *szórendfüggetlenség*.

2.3. BUSZI • A nyelvi jelenségek vizsgálatában kiemelten értékesek a beszélt nyelvi korpuszok. Ezek szöveganyaga eredetileg nem írott szöveg, hanem valódi szóbeli megnyilatkozások összessége, melyet utólag jegyeznek le meghatározott formában. Az ilyen korpuszok készítése jóval nagyobb erőfeszítést igényel, mint az írott nyelvi korpuszoké. A *Budapesti Szociolingvisztikai Interjú* felvételei a

2. ábra • A *hideg* és *hátán* bővítmények megadási módja, és az eredményül kapott, jellegzetesen e két bővítménnyel járó igék a Mazsola felületén.

1980-as évek végén készültek, az annotált, lekérdezővel ellátott korpusz 2012-ben vált elérhetővé. A korpusz szöveganyaga 270 000 szó.

A BUSZI-korpusz nagyon gazdag annotációval bír a a beszélt nyelvi jelenségek tekintetében, az annotációban rejlő nyelvi tudás a kereső segítségével minden részletében feltárható *(3. ábra)*.

A leggyakoribb *l*-kiesés mellett a magyarban a *d*-kiesés is sűrűn előfordul. Ha megvizsgáljuk a *majdnem* szó előfordulásait a BUSZI-korpuszban, kimondhatjuk, hogy beszédben majnem mindig kiesik a *d*.

Ha egy jelenség egy szón belül kétszer (többször) fordul elő, akkor a BUSZI korpusz keresője képes arra, hogy ezt két külön találatként jelenítse meg, azaz valójában itt az alapegység a szónál kisebb: ezt a keresőt nevezhetjük hangalapúnak.

A BUSZI-korpusz újdonsága annotációs szinten a számos, részletesen annotált, keresheto *beszélt nyelvi jelenség*. Lekérdezőszinten

…bizonyos dógokban ☐ mmm tát, hogy öööö lustább annál, mint amilyennek elképzeltem…

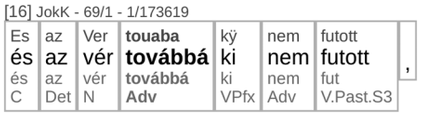
3. ábra • A BUSZI-ban annotált jelenségek illusztrációja. E rövid részletben számos beszélt nyelvi jelenség megtalálható, ezek a korpusz annotációjában mind explicit módon megjelennek. A négyzet szünetet jelöl, a *mmm* és az *öööö* hezitációt. Az annotáció tartalmazza a *tát* szónak a regularizált *tehát* alakját. A *dógokban* szó esetében szintén tudjuk a regularizált alakot *(dolgokban)*, és a szótövet *(dolog)*. Ezenkívül tudjuk azt is, hogy itt egy *l*-kiesés jelenséggel van dolgunk, mely mássalhangzó előtti pozícióban történt, s ami ún. *pótlónyúlással* (a kiesést megelőző hosszú *ó*-val) párosul.

is van egy új funkció: a táblázatként megjelenített *kétdimenziós*, alkorpusz szerinti gyakorisági eloszlás, mely egyrészt az adatközlők foglalkozása, másrészt az interjútémák szerint prezentálja az előfordulási számokat.

2.4. Ómagyar korpusz • Az eddigi három korpusz mind a mai magyar nyelv kutatását célozza. A 2013 óta elérhető *Ómagyar korpusz* (Simon – Sass, 2012) ezzel szemben egy jóval korábbi nyelvváltozatnak megfelelő régi magyar szövegeket tartalmaz. Jelentős eredmény, hogy az összes ómagyar korból származó (1526 előtti) magyar nyelvű kódex szövegét felöleli, ez összességében kétmillió szónyi anyagot jelent. Fontos, hogy a nagyon heterogén szöveganyag, a számos különleges ékezetes és egyéb karakter (*ȳ, ȳ* …) az Unicode kódrendszer használatával egységes formában és karakterkódolásban van meg. Betűhű átiratban rendelkezésre áll a teljes korpusz, bizonyos részek annotációja az esetleges helyesírási változatok egységesítését adó normalizálást, illetve morfológiai elemzést is magában foglal. Az ómagyar szövegek automatikus morfológiai elemzéséhez szükség volt a magyar morfológiai elemző jelentős átalakítására, mind a szókincs, mind a korabeli toldalékrendszer kezelésére alkalmassá kellett tenni.

Példa: A morfológiai elemzésnek köszönhetően vizsgálhatók például különféle szórendre vonatkozó kérdések. A korpuszadatok alátámasztják többek között azt a hipotézist, hogy a mai magyarban meglévő fordított tagadó szórend (például: *nem futott ki*) helyett az ómagyarban gyakran egyenes szórendet használtak (például: *ki nem futott*).

Az Ómagyar korpusz lekérdezőjének új funkciója a *párhuzamos megjelenítés*. Ez azt jelenti, hogy egy szövegnek a különféle szintjei párhuzamosan egymás alá rendezve jeleníthetők meg *(4. ábra)*.



4. ábra • Párhuzamos megjelenítés az Ómagyar Korpuszban. Egymás alatt látható a példaszöveg betűhű és normalizált változata, majd a szótövek és a morfológiai kódok következnek.

2.5. MNSZ2 • 2014-ben nyílt meg az MNSZ jelentősen bővített, felújított annotációval és lekérdezőfelülettel ellátott második változata (Váradi – Oravecz, 2014; Oravecz el al., 2015). Az új, pontosabb morfológiai elemzés kiterjed a szavakban található összes elem (így az *összetett szavak* tagjai, a *képzők*) azonosítására, ezenkívül az annotáció információt tartalmaz az egyes szóalakokról mint hangsorokról, lehetővé téve a fonológiai jellemzők (például: zöngés–zöngétlen) alapján való keresést. A használt modern korpuszkezelő rendszernek köszönhetően nagyon sok újdonság van a lekérdező funkcionalitásában. *Kollokációvizsgálat* segítségével feltérképezhetjük egy szó környezetében legjellemzőbb módon előforduló adott tulajdonságú egyéb szavakat, ezenkívül különféle *gyakorisági listákat* készíthetünk, vagy a kapott eredményt újabb lekérdezéssel tovább *szűrhetjük*. Az MNSZ első változatával ellentétben mindenféle korlátozás nélkül megkaphatjuk az adott lekérdezésre vonatkozó összes találatot.

Példák: az MNSZ2 felületén nem csupán a találati szavakból, hanem a találati szavakkal valamilyen viszonyban álló szavakból is készíthetünk gyakorisági listát *(5. ábra)*. Ez a funkció hasznos lehet szótárkészítés vagy stílusvizsgálatok során. Érdekes lehetőség, hogy az MNSZ2-ben a szóvégekre vonatkozó feltételek megadásával rímeket kereshetünk, s a kapott eredményből akár „verseket” is írhatunk.

5. ábra • Egy gyakorisági lista az MNSZ2-ből, mely a *légbőlkapott* melléknevet közvetlenül követő szavakból készült.

Az *es* végű mellékneveket követő *ek* végű többes számú főnevekre vonatkozó lekérdezés eredményéből állítható össze a következő „vers”: *kedves hölgyek / verses könyvek / szerves mérgek / csendes könnyek*. Vagy *si* és *ás* esetén ez: *gyimesi prímás / falusi zsongás / havasi tisztás / jézusi mondás*. Az MNSZ2-ben meglévő hangalapú fonológiai reprezentáció lehetővé teszi, hogy adott feltételeknek megfelelő hangsorból álló szavakra keressünk, és így például vizsgálhassuk a vegyes hangrendű szavak végén megjelenő *-ban/-ben* rag viselkedését. A szükséges lekérdezések lefuttatása és értékelése után azt tapasztaljuk, hogy míg a *kastély, szomszéd* szavakhoz kizárólag *-ban* járul, a *farmer, gundel* toldalékaként a *-ban/-ben* egyenlő arányban szerepel. *Hamlet* és *dzsungel* esetén főként *-ben*-t találunk, a *garden, tandem* pedig csak *-ben* ragalakkal fordul elő.

A fenti korpuszoknál idézett példákhoz hasonlóan e példák is alkalmasak arra, hogy a közoktatás keretein belül bemutassuk őket. Felsős vagy gimnáziumi magyarórán, fakultáción helyet kaphatnak az ilyen, éppen a nyelvészeti szövegkeresők segítségével megvalósítható egyszerűbb nyelvi vizsgálatok.

2.6. Összefoglalás • Az alábbi 1. táblázatban foglaljuk össze az eddigiekben említett annotációs, tartalmi, nyelvi funkciókat, illetve lekérdezési, formai, keresési funkciókat.

A korpuszok és korpuszlekérdezők fent áttekintett fejlődését három feltétel megléte tette lehetővé. Az egyre nagyobb teljesítményű számítógépeknek köszönhető az egyre nagyobb méretű korpuszok kezelése, a nagy méret mellett is gyors lekérdezés; az egyre jobb, fejlődő elemzőeszközöknek a korpuszok annotációjában kódolt egyre részletesebb és szofisztikáltabb nyelvi tudás; az egyre jobb, fejlődő korpuszkezelő rendszereknek pedig az újabb és újabb lekérdezőfunkciók.

Bár vannak szinte minden esetben meglévő, alapvető, sztenderdnek tekinthető funkciók – az *1. táblázat* mindkét kategóriájában az első kettő –, az is szembeötlő, hogy milyen változatos, eltérő funkciókkal bíró, eltérő célokra használható korpuszok születtek. A használt korpuszkezelő rendszer sem azonos, mert esetenként érdemesnek tűnt újabbra váltani a jobb funkcionalitás miatt.

Hasznos lenne összegyűjteni, és valamilyen módon egységesíteni, egyben kezelni, minél kiterjedtebb annotációval és lekérdezési funkciókkal ellátni az összes hazai korpuszt. Erről a jövőbeli kívánatos fejlődési irányról szól a következő rész.

	MNSZ1	Mazsola	BUSZI	Ómagyar	MNSZ2
alapegység					
	szó	tagmondat	hang	szó	szó
annotációs funkciók					
szótő	✓	✓	✓	✓	✓
morfológia	✓		✓	✓	✓
szórendfüggetlenség		✓			
beszélt nyelvi jelenségek			✓		
spec. karakterek, ómagyar morfológia				✓	
összetett szavak, képzők, fonológia					✓
lekérdezőfunkciók					
konkordancia	✓		✓	✓	✓
szókapcsolatra keresés	✓	✓	✓	✓	✓
kétdimenziós gyakorisági eloszlás			✓		
párhuzamos megjelenítés				✓	
szűrés, gyakorisági listák, kollokáció					✓

1. táblázat • A bemutatott korpuszok és korpuszlekérdezők jellemzői

3. Nemzeti Korpuszportál

A *Nemzeti Korpuszportál* (NKP) kezdeményezés célja, hogy együtt, egy helyen megtalálható legyen minden magyar nyelvű, online lekérdezhető korpusz, és ugyanott megtalálható legyen minden elérhető lekérdező funkció.

A létrejövő online portál feladata hármas. Egyrészt nagyon fontos, hogy egy ponton hozzáférést biztosítva a hozzáférhető magyar nyelvű korpuszokhoz, népszerűsítse a korpuszokat és a korpuszhasználatot a szakma és a nagyközönség körében. Egyaránt szeretnénk megszólítani a nyelvészeket, irodalmárokat, bölcsészeket, magyartanárokat, diákokat, és minden nyelvi kérdések iránt érdeklődő embert. Másrészt szeretnénk bemutatni, reklámozni a korpuszhasználóknak a portálon elérhető többi korpuszt. Harmadrészt pedig szeretnénk elérni, hogy a magyar nyelvű korpuszokkal foglalkozó műhelyek szorosabb szakmai kapcsolatba kerüljenek egymással, egymás megoldásait tanulmányozhassák és alkalmazhassák saját korpuszaikra, és könnyebben alakíthassanak ki közös projekteket, együttműködéseket.

A következőkben bemutatjuk az NKP négy „szintjét". Ezek a jövőben elérendő egyre magasabb fejlettségi szinteket jelentik.

Nulladik szinten megpróbálunk minél több korpuszt felkeresni, és e korpuszokhoz mindössze néhány alapvető adatot gyűjtünk össze: név, online lekérdező linkje, kapcsolat (a korpusz gazdájának e-mail címe). Már ezzel a nulladik szintű portállal lényegében megvalósíthatók a fent részletezett feladatok.

Első szinten az *1. táblázathoz* hasonló funkciótáblázat készül majd, mely információt ad arról, hogy melyik korpusz milyen annotációt tartalmaz, és milyen keresőfunkciói érhetők el. Ezen kívül kiegészítő adatok (létrehozás ideje, leírás, egy illusztratív példa, angol felület stb.) felvétele is hasznosnak tűnik.

A második szint nagy ugrást jelent az elsőhöz képest. Itt az a cél, hogy az összes, bármelyik korpusznál elérhető lekérdezőfunkció elérhető legyen az összes korpuszra, másképp: hogy a funkciótáblázat második felét lehetőleg teljesen kitöltsük pipákkal. Ehhez az szükséges, hogy absztraháljuk a lekérdezőfunkciókat, azaz olyan rendszert alakítsunk ki, ahol az egyes lekérdezőfunkciók önálló objektumokként, entitásokként kezelhetők. Ha ez megvan, akkor a lekérdezőfunkciók (gyakorisági lista, szűrés stb.) szabadon hozzárendelhetők az egyes korpuszokhoz, esetleges korlátot csak az jelenthet, ha az adott korpusz annotációja egy lekérdezőfunkciót nem enged meg. Valószínűleg az MNSZ2-ben használt, gazdag funkcionalitással bíró korpuszkezelő rendszerből kiindulva lehet az NKP második szintjét legkönnyebben megvalósítani. Második szinten egy saját korpusz közzétételéhez elegendő lesz az annotált szöveganyagot (az „XML-t") elkészíteni, a kívánt lekérdezőfunkciókat csupán (kézzel vagy esetleg automatikusan) hozzá kell rendelni a portálon.

A harmadik, egyben végső szint újabb nagy ugrást jelent. A cél itt az, hogy az összes annotációs funkció is elérhetővé váljon tetszőleges korpuszra. Ehhez a magyar nyelvű szövegek különféle annotációit előállító, nyelvi tudással bíró elemzőeszközök gyűjtése és közzététele is szükséges a portálon. Ezzel az egyes annotációs funkciók is önálló egységekként hozzáférhetővé válnak, és szabadon hozzárendelhetők az egyes korpuszokhoz, azaz a funkciótáblázat első felében is elszaporodhatnak a pipák. Harmadik szinten egy korpusz közzétételéhez elegendő lesz pusztán az elemzetlen szöveganyagot (a „TXT-t") összeállítani, a kívánt annotációs és lekérdező-

funkciókat a portálon rendelhetjük majd hozzá. Jelenleg is folynak munkálatok hasznos magyar nyelvi elemzőeszközök tökéletesítésére és közzétételére.

A Nemzeti Korpuszportál működése tehát a távoli jövőben így nézhetne ki:

1. összeállítjuk a szöveganyagot (például adott irodalmi művekből vagy egy témában gyűjtött Facebook-bejegyzésekből);
2. sima szövegként feltöltjük az NKP-re, megmondjuk, hogy szükségünk van mondatra bontásra, szótövesítésre, morfológiai elemzésre (3. szint);
3. illetve konkordanciára, gyakorisági listára, kollokációkeresésre (2. szint);
4. lefutnak az elemzések, elkészül az annotált korpusz, hozzárendelődnek a lekérdezőfunkciók, automatikusan előáll a funkciótáblázat megfelelő oszlopa (1. szint);
5. valamint az új bejegyzés az NKP nyitólapján (0. szint).

Ezzel készen van, lehet kutatni, lekérdezni az új korpuszt.

A Nemzeti Korpuszportál 2015 novemberében elindult, jelenleg legelső nulladik szintű verziójában érhető el az URL1 internetes címen. A fentiekben áttekintett korpuszokat tartalmazza, illetve a lista máris bővült több korpusszal.

Fontos cél, hogy minél több magyar nyelvű korpusz hozzáférhető legyen a portálon, ennek érdekében mindenkit, akinek van megfelelő korpusz a birtokában, felhívunk a csatlakozásra. A portálra az általános célú korpuszok mellett bármilyen speciális célú korpuszt is szívesen fogadunk. A csatlakozás feltételei részletesen olvashatók a honlapon. Fontos feltétel a magyar nyelvű szöveganyag, a keresett jelenség minden előfordulását külön találatként megjelenítő szóalapú (esetleg hang-, vagy tagmondatalapú) online lekérdező, és egy korpuszgazda e-mail címmel, akihez szükség esetén fordulni lehet. Ezen kívül minden tagot kérünk, hogy a többi korpusz népszerűsítése, reklámozása érdekében lehetőleg helyezzen el egy NKP-ra mutató linket a saját korpusz oldalán.

A bemutatott korpuszokhoz nem közöltünk internetes elérhetőségeket, mivel mindegyik megtalálható az NKP-n az URL1 címen. Próbálják ki, használják a korpuszokat, tanulmányozzák a fent említett példákat, kérdés esetén pedig bátran forduljanak a korpuszgazdákhoz vagy a cikk szerzőjéhez!

Kulcsszavak: *korpusz, korpuszlekérdező, Nemzeti Korpuszportál, annotáció, konkordancia, szóosztály, nyelvi adat, nyelvészet*

IRODALOM

Kilgarriff, Adam (2007): Googleology Is Bad Science. *Computational Linguistics.* 33, 1, 147–151. DOI: 10.1162/coli.2007.33.1.147 • http://dl.acm.org/citation.cfm?id=1245144

Kiss Gábor (1998): *Magyar szókincstár.* Tinta, Budapest

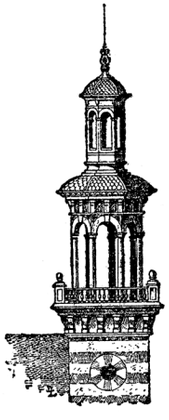
Moretti, Franco (2013): *Distant Reading.* Verso, London

Oravecz Csaba – Sass B. – Váradi T. (2015): Mennyiségből minőséget. Nyelvtechnológiai kihívások és tanulságok az MNSz új változatának elkészítésében. In: Tanács Attila – Varga V. – Vincze V. (szerk.): *XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2015).* JATEPress, Szeged • http://rgai.inf.u-szeged.hu/mszny2015/files/MSZNY2015_press_B5_PQ.pdf

Sass Bálint (2009): „Mazsola" – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradi Tamás (szerk.): *Válogatás az I. Alkalmazott Nyelvészeti Doktoranduszok Konferencia előadásaiból.* MTA Nyelvtudományi Intézet, Budapest • http://www.nytud.hu/alknyelvdok07/proceedings07/Sass.pdf

Simon Eszter – Sass Bálint (2012): Nyelvtechnológia és kulturális örökség, avagy korpuszépítés ómagyar kódexekből. In: Prószéky Gábor – Váradi Tamás (szerk.): *Általános Nyelvészeti Tanulmányok.* XXIV, 243–264. • http://omagyarkorpusz.nytud.hu/

documents/papers/chapters/simonsass-chapter_-_nyelvtech_es_kult_orokseg.pdf
Váradi Tamás (2002): The Hungarian National Corpus. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002).* • http://www.lrec-conf.org/proceedings/lrec2002/
Váradi Tamás – Oravecz Csaba (2014): A Magyar Nemzeti Szövegtár egymilliárd szavas új változata. *Magyar Tudomány.* 175, 9, 1054–1062. • http://www.matud.iif.hu/2014/09/05.htm
URL1: Nemzeti Korpuszportál http://corpus.nytud.hu/nkp



A DIGITÁLIS KORSZAK VÍVMÁNYAINAK HASZNOSULÁSA A LEXIKOGRÁFIÁBAN: A NAGYSZÓTÁRI PROJEKT INFORMATIKAI FEJLESZTÉSEIRŐL

Simon László

tudományos segédmunkatárs,
MTA Nyelvtudományi Intézet
simon.laszlo@nytud.mta.hu

Bár *A magyar nyelv nagyszótára (Nagyszótár)* eredetileg (a múlt század közepén) papíralapú szótárként lett kitalálva és megtervezve, amikor a jelenlegi munkálatok elkezdődtek, valójában már számítógépes adatbázist, adatbázisokat építettünk, amelyeknek egyik lehetséges megjelenési formája, kimenete a szótár nyomtatott változata: a jól strukturált adatbázisokban tárolt információk egymással való összekapcsolása nyilvánvalóan több lehetőséget rejt magában annál, mint amit egy klasszikus nyomtatott kötet(sorozat) nyújtani tud.

A *Nagyszótár*[1] egynyelvű, elsősorban értelmező típusú szótár, történeti vonatkozásokkal: azaz „nem csupán a mai nyelvállapot szókészletét tükrözi és elemzi, hanem történeti dimenziót is érvényesít" (Ittzés, 2011). A húszkötetesre és mintegy 110 ezer címszavasra tervezett sorozat eddig megjelent darabjai egy olyan szövegadatbázison alapultak, amely 1772 és 2000 közötti szemelvényeket tartalmazott, az előkészületben lévő hatodik és a hetedik kötet azonban már a 2001 és 2010 közötti időszakból is felvesz adatokat. A *Magyar történeti szövegtárban* ugyanis a nyelvújítást megelőző időszakra jellemző, a korai forrásainkban bőséggel megtalálható archaizmusok, elavult szavak vagy jelentések mellett ma már a XXI. század elejének szóhasználatát, az egyes szaknyelvek terminológiájának változásait, újításait illusztráló adatok is a rendelkezésünkre állnak. Az alábbiakban vázlatosan ismertetjük azokat az informatikai újításokat, megoldásokat, amelyek a szótárírás munkáját nagyban elősegítik.

[1] A hivatalos, nyelvészeti írásokban alkalmazott rövidítés az Nszt., de használatától a közérthetőség végett itt most eltekintünk.

A történeti korpusz bővítése karakterfelismertetéssel

A nagyszótári munkálatok szempontjából a legfontosabbnak számító szövegtár (korpusz) és az informatika kapcsolatáról a szótár 1. kötetében, az ún. *apparátuskötetben* a következőket olvashatjuk: „…az 1980-as években elektronikus formában hozzáférhető szövegek még nem álltak rendelkezésre, a *Nagyszótár* előkészítésén fáradozó munkacsoport út-