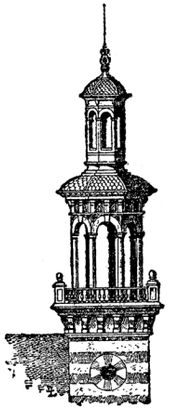
documents/papers/chapters/simonsass-chapter_-_nyelvtech_es_kult_orokseg.pdf
Váradi Tamás (2002): The Hungarian National Corpus. In: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002).* • http://www.lrec-conf.org/proceedings/lrec2002/
Váradi Tamás – Oravecz 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-

törő munkát végzett e téren. A kézi szövegrögzítés a mai lehetőségekhez képest rendkívül kezdetleges Commodore számítógépekkel kezdődött, az ékezetes magyar magánhangzókat, illetve a történeti karaktereket ekkor még csak betűk és számok kombinációjával kialakított kódokkal tudtuk megjeleníteni. A szövegek tárolására alkalmas, nagy kapacitású számítógép akkoriban csak a Számítástechnikai és Automatizálási Kutatóintézetben állt rendelkezésünkre, a begépelt adatokat oda kellett szállítani. Az 1985-től 2005-ig tartó időszakban a korpuszépítés során – fokozatosan egyre modernebb eszközökkel és a munkálatokat mindinkább kiszolgáló szoftveres háttérrel dolgozva – létrehoztunk egy több mint 27 millió szövegszavas elektronikus szövegbázist." (Csengery, 2006)

Ez az adatbázis, a *Magyar történeti szövegtár* ún. *xml*-állományokon alapszik: a szövegek „jólformázottságát" az egyes alkotóelemek – cím, bekezdés, idézet stb. – felcímkézésének köszönhetjük (az *xml*-formátumról részletesebben alább, a szócikkírás számítógépes hátteréről szóló részben ejtünk szót). Ám amíg az 1772 és 2000 közötti szövegeket reprezentáló, 2500 szerzőtől származó 22 000 szemelvény rögzítése a címkék közé történő begépeléssel valósult meg, addig a szótári osztályon 2014–2015-ben folyó korpuszbővítés során (a bővítés a *Magyar történeti szövegtár* 2001 és 2010 között keletkezett szövegekkel való kiegészítését jelenti, mintegy 2,9 millió szövegszó terjedelemben) az optikai karakterfelismertetésnek jutott a főszerep. Mivel az ezt a feladatot elvégző szoftver (esetünkben az *ABBYY FineReader*) számára a „nyersanyag", a kiindulópont valamilyen képformátumú állomány, azokról a nyomtatványokról (könyv- és újságoldalakról), amelyek az általunk a korpuszba emelni kívánt szövegeket tartalmazták, fotódokumentációt kellett készítenünk. Ez néhány száz oldal esetében ténylegesen digitális fényképezőgéppel való képkészítést jelentett, a legtöbbször azonban szkennerrel (lapolvasóval) történő rögzítést.

Mihelyt a képfájlok rendelkezésünkre álltak, a karakterfelismertető programmal megkezdődhetett a szövegfájlok előállítása, amelyek a szoftver „érzékenysége" és az általa használt szótárak fejlettsége következtében egy átlagos dokumentumoldal adattartalmát – ideértve a formázásokat is – közel százszázalékos pontossággal adták vissza. Az *ABBYY FineReader* azokat a szöveghelyeket, ahol a karakter beazonosítását bizonytalannak ítéli, külön meg is jelöli, így a kortárs regényekből és verseskötetekből, napi- és hetilapokból, valamint egyetemi és középiskolai tankönyvekből, monográfiákból, naplókból és szociográfiákból, továbbá szakfolyóiratokból, riport- és interjúkötetekből származó, a karakterfelismertetés módszerével létrehozott digitális tartalmakat már nem kellett a hagyományos módon összeolvasni, azaz a nyomtatott változattal betűről betűre összevetni. Elegendő volt csupán a karakterfelismertető jelöléseit, illetve a Word helyesírás- és nyelvhelyesség-ellenőrzőjének javaslatait elbírálni, és azok figyelembevételével a javításokat elvégezni.

Az *ABBYY FineReader* segítségével előállított szemelvényeket a szükséges korrekciót követően konvertáltuk *xml*-formátumúvá, a szöveg elemei e folyamat eredményeként kapták meg a korpuszfájlok dokumentumtípus-definíciója szerinti címkéjüket.

Amikor meghoztuk azt a döntést, hogy a korpuszbővítés során felhasznált forrásokat már csak elektronikusan – képállományok vagy pdf-fájlok formájában – tároljuk, az talán

már egy szükségszerű elmozdulás volt a papíralapúságtól a digitális világ felé.

A nagyszótári cédulák digitalizálása

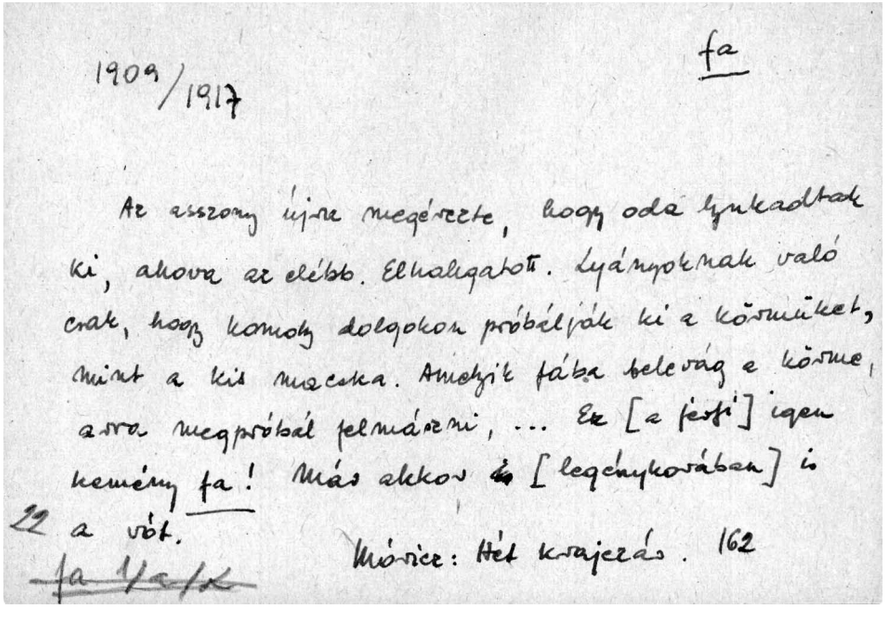
A *Nagyszótár* eredeti papíralapú forrása, a mintegy hatmillió darabos *archivális cédulagyűjtemény* digitalizálására az elmúlt tizenöt évben kétszer is kísérletet tettünk, de a feldolgozottság még mindig csak 25 százalékos. Ezek közül a rendszerint pontosan A6-os, 14,8 × 10,5 centiméter méretű lapok közül a legkorábbiakat a 19. század végén készítették, de a számukat a szótári munkálatok során ma is gyarapítjuk, mivel ún. pótcédulákat százával állítunk elő. A cédulák adattartalmának legfontosabb eleme mindig az idézet, amelynek hossza a néhány szavastól a több bekezdésnyiig változik, de megtalálható rajtuk az idézett mű keletkezésének éve és természetesen a címszó is. A cédulára kiírt szöveg szerzőjének a neve, a felhasznált mű címe általában a lap jobb alsó sarkában olvasható.

Az Akadémiai Nagyszótár történetének vázlata (1898–1952) c. tanulmányban a cédulázás korai időszakáról az alábbiakat tudhatjuk meg: „…arra utasították a gyűjtőket, hogy a nyelv egész szókincsére legyenek tekintettel. Az utasítás szerint még a meghonosodott idegen szót is gyűjteni kellett. Mellőzték azonban a tréfás szóhasználatot, az új, el nem terjedt neologizmusokat, a mindennapi szavakat (pl. *víz, ház, kenyér, asztal*), ez utóbbira csak akkor voltak tekintettel, ha egy-egy ilyen szó új jelentésárnyalatot hordozott. Óvatosan kellett gyűjteni a szólásokat. Tilos volt azonban mondattani jelenségeket és tulajdonneveket cédulázni." (R. Hutás, 1973)

A későbbi korszakokban készült cédulákra is igen jellemző, hogy rendszerint olyan szövegrészletek hordozói, amelyekben az adott szó egy különleges, nem hétköznapi jelentésében fordul elő, ebből következően a szótári munka során sokszor azért vesszük kézbe őket, hogy egy ritkább árnyalat illusztrálásához találjunk példát. A két digitalizálási projekt eredményeként jelenleg mintegy másfél millió szkennelt cédula képfájlja áll rendelkezésünkre. A cédulagyűjteménynek, ennek az unikális, a műfajából adódóan egyetlen példányban létező és igen sérülékeny forrásnak a szkennelésekor a legfontosabb célkitűzésünk a biztonsági másolat létrehozása volt. Az adattartalmuk szövegkereső programok számára való elérhetővé tétele egyelőre túl ambiciózus elképzelésnek tűnik, hiszen a cédulák túlnyomó része kézírásos, így a karakter felismertetésekor az egyes betűk azonosításához több tucatnyi különböző kéz írásképét kellene elemezni.

Bár a cédulák a dobozokban egymást szoros betűrendben követik, így egy-egy szó anyaga könnyen megtalálható, a képfájlok léte már önmagában előrelépést jelent abban a tekintetben, hogy akárhány másolat készíthető belőlük, és egy egyszerű képszerkesztővel böngészhetőek. Rövid távon csak annyit szeretnénk elérni, hogy valamilyen kulcsszavazási technikával legalább odáig eljussunk, hogy a fájlokhoz hozzárendeljük azt az információt, hogy melyik címszó adatát hordozzák, illetve annak a műnek az azonosítóját is, amelyből a cédulán szereplő idézet való.

A cédulaanyagnak a tárolódobozokba való elhelyezésekor, illetve betűrendbe sorolásakor begépelt szójegyzéket már évekkel ezelőtt elérhetővé tettük: az *nszt.nytud.hu* webcímen a mintegy hétezer oldalas dokumentum minden egyes lapja megtalálható. A megtekintéshez a fenti címen *Az archivális cédulagyűjtemény címszójegyzéke* menüpontra kell kattintanunk, ahol kiválaszthatjuk az ábécé minket érdeklő betűjét.



1. kép • A hatmilliós archiválisz cédulagyűjtemény egy darabja

A szócikkírás számítógépes háttere

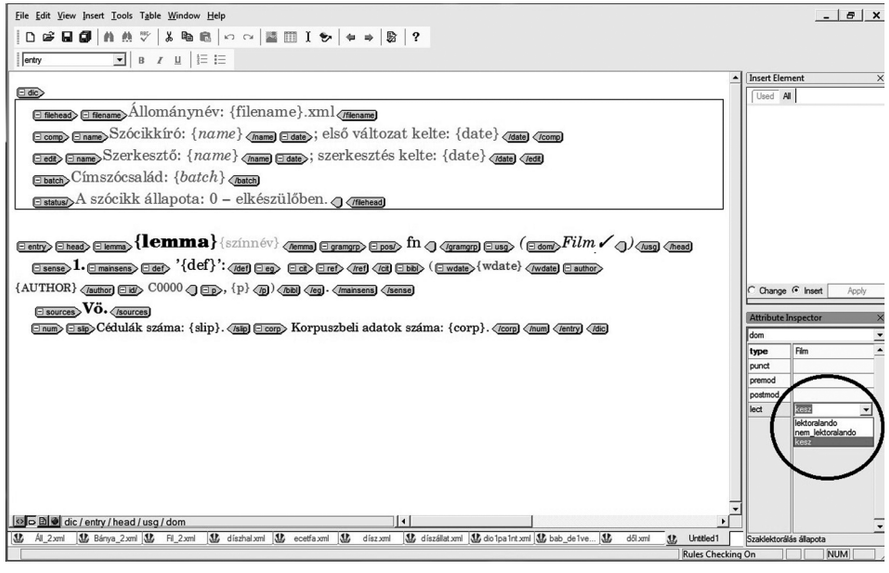
A nagyszótári szócikkek már több mint egy évtizede *xml*-formátumban íródnak és tárolódnak, aminek következtében már csírájukban is egy adatbázis elemeként léteznek. Az *xml* az angol *Extensible Markup Language*, azaz a „kiterjeszthető jelölőnyelv" kifejezés rövidítése, a „jelölő" pedig annyit tesz, hogy a szócikk szövegének, struktúrájának minden egyes fontosabb részlete mintegy fel van címkézve. Más címke (ún. *tag*) azonosítja be például a példamondat kezdetét és végét, és megint egy másik az idézet keletkezésének idejét. A szócikkeket tehát nem egy hétköznapi szövegszerkesztő, hanem az *xml*-ek szerkesztésére fejlesztett program, az *XMetaL* segítségével írjuk, amely lehetőséget ad arra, hogy a szövegnek a képernyőn/nyomtatásban való megjelenítése az egyes címkékhez hozzárendelt formázásokkal történjék meg. Az *XMetaL*-ban a címkék ún. attribútumokkal is elláthatóak, ami a formalizálható megoldások esetében sokszor megkönnyíti és felgyorsítja az egyes részfeladatok elvégzését. Ha vetünk egy gyors pillantást a lexikai minősítések használatát rögzítő szabályokra, láthatjuk, hogy ezeknek az egyszerűnek tűnő megjegyzéseknek a szócikkben való elhelyezésekor milyen sok apróságra kell odafigyelni.

2. kép • Egy-egy dobozban öt-hatezer cédula lapul meg

„Több lexikai minősítés meghatározott sorrendben kerül a szócikkbe. Ha lexikai minősítések érvényessége időben határolódik el egymástól, törtjellel (/) választjuk el egymástól a korábbi szinkróniára, illetve a mai használatra érvényes jelölést. […] Ha egy minősítés a teljes korszakra vonatkozik, a törtjelet elhagyjuk, például: **abbiz** *(nyj),* **ablakszem** *(ritk).* […] Ha több olyan minősítést tüntetünk fel, amely a teljes korszakra érvényes, ezek közös zárójelben kerülnek a szócikkbe. Ha a minősítések az adatok összességére vonatkoznak, akkor egymástól vesszővel elválasztva vesszük föl őket, például: *(nyj, durva).* […] Ha több lexikai minősítés törtjel nélkül kerül közös zárójelbe, a következő sorrendben vesszük föl őket: 1. helyen a szó időbeli elterjedtségére, illetve gyakoriságára utaló *(rég)* és *(ritk)* minősítés áll; a 2. helyre a területi vagy csoport- és rétegnyelvi minősítések kerülnek, ezen belül 2/1. a területi használatra utaló *(nyj),* 2/2. a szaknyelvi minősítés, például: *(Áll), (Műv),* 2/3. a rétegnyelvi, például: *(irod), (biz), (hiv);* a 3. helyen a stíluséréket jelölő rövidítések állnak, például: *(tréf), (pejor).*" (Ittzés, 2006)

Az ezekhez a minősítésekhez rendelhető attribútumok rendszere biztosítja többek között azt is, hogy a különböző információk a szabályzatban előírt, kötött sorrendben kövessék egymást. A szócikkek szerkezetét rögzítő *dokumentumtípus-definíció* egyik legfrissebb módosítása a munkálatok részét képező szaklektoráltatást könnyíti meg. Ennek a többlépcsős folyamatnak a keretében először a szaknyelvi minősítéssel ellátott szócikkeket el kell juttatnunk az adott tudományterületet behatóan ismerő lektornak, majd a visszakapott szakvéleményeket nyilvántartásba kell vennünk. Csak ezután következhet az a fázis, amelynek során a szócikk szerzője és szerkesztője tanulmányozza a szaklektor észrevételeit, és dönt a javasolt módosításokról. Ám ők akár változtatnak a szócikken, akár nem, azt mindenképp adminisztráljuk, hogy a szerzők a lektor álláspontját megismerték, és annak figyelembevételével dolgozták át vagy hagyták érintetlenül az értelmezést, a példákat stb. A legújabb fejlesztések eredményeként a jelenleg szerkesztés alatt álló hatodik kötet mintegy 4200 szócikke közel 2200 szaknyelvi minősítésének mindegyike tartalmazza azt az információt, hogy (1) még lektorálás előtt áll; (2) nem kell lektoráltatni; (3) már lektorált. Nem kell külön adminisztrálni, hogy a lektori jelentések maradéktalanul fel lettek-e dolgozva, hiszen a szaklektorálási folyamat lezárultát jelző attribútum minden egyes *Orvos, Ipar* vagy *Növ* stb. minősítés esetében beállítható/beállítandó. Mindennek következtében azt is elértük, hogy a szócikkállomány alapértelmezett tárhelyén futtatható egy olyan segédprogram, amely egyszerűen kigyűjti és csoportosítja az aktuálisan még lektorokhoz küldendő jelentésegységeket.

A nagyszótári munkálatok során nap mint nap módosuló szócikkállomány, a tevékenységünket támogató egyéb segédfájlok (útmutatók, listák stb.), valamint a *Magyar történeti szövegtár* és a forrásjegyzékünk *xml*-jeinek naplózását (logolását), a változások követését biztosítandó 2015 szeptemberében az SVN



3. kép • Képernyőkép a XMetaL ún. tages nézetéről: a szócikkstuktúra egyes részeit ilyen címkék közé helyezzük el. Jobbra a szaknyelvi minősítések lektoráltsági állapotának kódolását lehetővé tevő attribútumlista látható.

(*Subversion*) bevezetése mellett döntöttünk. Ez a széles körben elterjedt, nyílt forráskódú verziókövető rendszer lehetővé teszi számunkra, hogy a munkacsoport által közösen használt, egy központi szerveren tárolt fájlok (egyúttal azok változatai, a módosítások tartalma, időpontja vagy akár szerzője stb.) internetkapcsolaton keresztül ma már bárhonnan elérhetőek, visszakereshetőek.[2]

Az SVN kliensprogramja (esetünkben a *TortoiseSVN*) gyorsan telepíthető a felhasználó számítógépére, és a regisztráció során beállított jogosultságokkal szabályozható a munkacsoport tagjainak az egyes könyvtárakra vonatkozó hozzáférése is. A munkavégzés nem igényel folyamatos internetkapcsolatot, hiszen a felhasználó saját gépén található fájlok csak a megfelelő parancs következtében szinkronizálódnak a szerveren tárolt változatokkal. Az *Update* parancs eredményeként minden fájlból a legfrissebb változatok töltődnek le, míg a *Commit* hatására a lokális állományok töltődnek fel a szerverre: minden egyes *Commit* egy külön elmentett, archivált verziót jelent. (A szótári osztályon szócikkek szerzői mindennap az *Update* futtatásával kezdik, és a *Commit* paranccsal fejezik be a munkavégzést.) A verziókövető rendszer egyik fontos előnye, hogy képes a több forrásból származó, párhuzamos módosítások összefésülésére, illetve, ha az sikertelen, akkor a konfliktusos állapot jelzésére is.

Látható, hogy a számítógépes rendszerekkel támogatott szótárírás nem csupán gyökeresen más, mint amilyenek a hajdani papíralapú munkálatok voltak, hanem a gépi háttér lehetőségeinek változása, bővülése folyamatosan javítja a *Nagyszótár* színvonalát is. Ráadásul 2015-re megteremtődött az a stabil informatikai háttér, amely hatékonyan segíti a nagyszótári munkálatokat. Sőt reményeink szerint ez év végére *A magyar nyelv nagyszótára* internetes elérhetőségét is sikerül biztosítanunk, miáltal a szócikkállományok minden érdeklődő számára kereshetővé válnak.

[2] A tárhelyet a MorphoLogic Kft. biztosítja számunkra, amiért ezúton is köszönetet mondunk, ahogy a szótári osztály munkaállomásain a rendszer „beüzemelését” végző Merényi Csabának is.

Kulcsszavak: *szócikk, xml, XMetaL, szövegadatbázis, karakterfelismertetés, képfájl*

IRODALOM

Csengery Kinga (2006): Az elektronikus korpusz. In: Ittzés Nóra (főszerk.): *A magyar nyelv nagyszótára 1. Segédletek*. MTA Nyelvtudományi Intézet, Budapest

R. Hutás Magdolna (1973): Az Akadémiai Nagyszótár történetének vázlata (1898–1952). *Nyelvtudományi Közlemények*. 75, 1, 447–465.

Ittzés Nóra (2006): Tájékoztató a szótár szerkesztési elveiről, szerkezetéről és használatának módjáról. In: Ittzés Nóra (főszerk.): *A magyar nyelv nagyszótára 1. Segédletek*. MTA Nyelvtudományi Intézet, Budapest

Ittzés Nóra (2011): *A magyar nyelv nagyszótárának* lexikográfiai koncepciója, *különös tekintettel a szemantika és a grammatika összefüggésére a szótárírásban* (doktori disszertáció). Szegedi Tudományegyetem, Szeged http://doktori.bibl.u-szeged.hu/1087/1/IttzesN-2011_disszertacio.pdf