KERESÉS KORPUSZBAN: A KIBŐVÍTETT MAGYAR TÖRTÉNETI SZÖVEGTÁR ÚJ KERESŐFELÜLETE

SASS BÁLINT
(MTA Nyelvtudományi Intézet)

Bevezetés – Nszt. és MTSz.

Az MTA Nyelvtudományi Intézetében folynak A magyar nyelv nagyszótára (Nszt.) (Ittzés 2011) munkálatai. Az I. kötet 2006-ban, az legutóbbi (VI.) kötet 2017-ben jelent meg, e kötetek az *a-ek*-ig terjedő címszavakat tartalmazzák. A tervek szerint a teljes szótár 20 kötetből fog állni, 110 ezer címszó fog helyet kapni benne.

A Nagyszótár történeti igényű, korpuszalapú magyar értelmező szótár. A korpuszalapúság azt jelenti, hogy a szótárkészítés során a lexikográfiai döntéseket nagyméretű szöveganyag vizsgálata alapján hozzák meg. A korpuszépítési munkálatok már 1985-től elkezdődtek, ennek során jött létre a Magyar történeti szövegtár (MTSz.) (Csengery 2006). Ez az a korpusz, amin a Nagyszótár elsősorban alapul.

Az MTSz. első változata 27 millió szövegszónyi anyagot ölelt fel az 1772–2000-ig terjedő időszakból. Ezt később kibővítették újabb 3 millió szövegszónyi anyaggal a 21. század első évtizedéből. A bővítés 2015-ben készült el. Az MTSz. tehát jelenleg 30 millió szövegszónyi anyagot tartalmaz az 1772–2010-ig terjedő majd 240 évből.

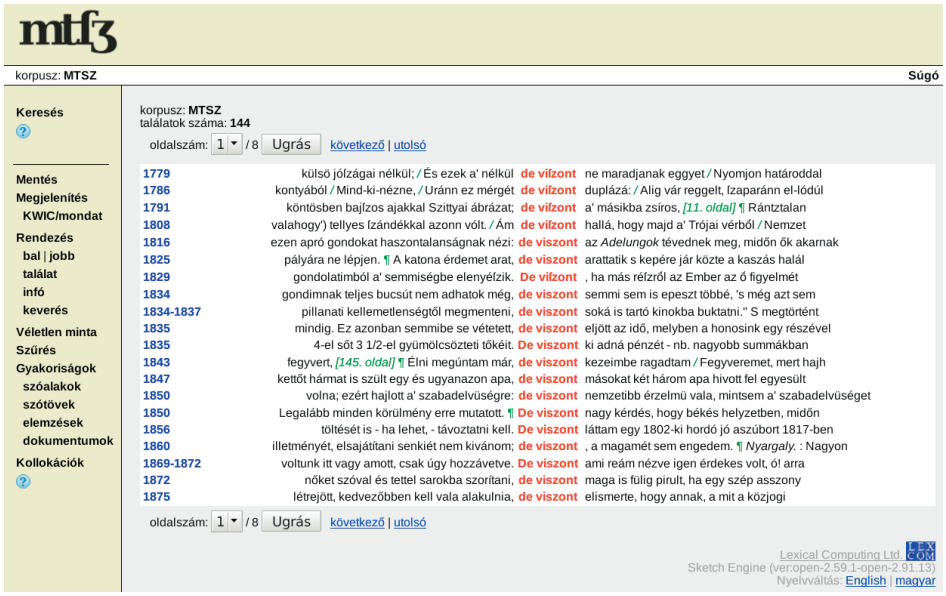
Fontos megjegyezni, hogy az MTSz. **elemzetlen** korpusz. Az elemzett korpuszokban az egyes szavakhoz különféle kiegészítő információk tartozhatnak, mint például szótő vagy morfológiai elemzés. Ilyen az MTSz.-ben nincs, azaz a szöveganyagra egyszerűen szóalakok sorozataként tekinthetünk.

A kibővített korpuszhoz 2016-ban új lekérdezőfelületet készítettünk, amely szabadon hozzáférhető kutatási célra és a nagyközönség számára is.

Az MTSz. használata, funkciói

Tekintsük át, hogy az Magyar történeti szövegtár új, http://clara.nytud.hu/mtsz címen elérhető lekérdezőfelülete milyen funkciókkal bír, milyen lehetőségeket kínál. Az 1. ábrán a *de viszont* kifejezésre lefuttatott egyszerű keresés eredménye látható.

A nyelvtörténeti kutatások újabb eredményei IX. Szerkesztette: Forgács Tamás – Németh Miklós – Sinkovics Balázs. Szeged: SZTE Magyar Nyelvészeti Tanszék. 267–277.



1. ábra:
A *de viszont* szókapcsolatra vonatkozó egyszerű keresés eredménye a Magyar történeti szövegtárban

A korpuszban 144 találat van. Az eredményben kis- és nagybetűvel egyaránt megkapjuk a találatokat, sőt a normál *s* mellett a hosszú szárú *ſ* is megjelenik. Ez annak köszönhető, hogy az Unicode karakterkódolási rendszer, melyet a többféle régi karaktert tartalmazó korpusz kódolására használtunk, a *ſ*-t is egyfajta *s*-nek tekinti, azaz betűrendbe sorolás szempontjából szerencsésen egy kalap alá veszi a kettőt. A szöveghez rendelt strukturális információt a zölddel megjelenő közbevetések tartalmazzák: ilyenek az oldalszámok, bekezdések, verssortörések. A találatokat időrendben látjuk, a nyelvi adat keletkezési ideje nyilván kiemelten fontos adat egy történeti szövegtár esetében. Megjegyezzük, hogy a kérdéses kifejezésre már 1779-ből – vagyis az MTSz. gyűjtési időintervallumának legelejéről – van adat, ami nem is meglepő a vonatkozó szakirodalom alapján (Schirm 2014a, 2014b, 2015).

A felület mögött működő **NoSkE** (NoSketchEngine) korpuszkezelő motor (Rychlý 2007) számos olyan funkcióval bír, amelyek az 1. ábrán közvetlenül nem látszanak. Ezeket érdemes a felületen élőben kipróbálni. A Nagyszótár igényei szerint kialakított részletes bibliográfiai adatok az adott találat előtt álló évszámra kattintva érhetők el, a találati szóra kattintva pedig nagyobb kontextusban vizsgálhatjuk az adott szövegrészt.

A bal oldali menüben a következő funkciókhoz férünk hozzá. Még a keresés elindítása előtt szűkíthetjük a szöveganyagot a teljes MTSz. egy **alkorpusz**ára. A kívánt alkorpuszt szerző, cím, oldalszám, vers/próza, szépirodalmi jelleg, regiszter és a megjelenés éve alapján határolhatjuk be.

További feldolgozásra **elment**hetjük az összes találatot. A **megjelenítés**nél beállíthatjuk a kontextus méretét, az egy oldalon megjelenő találatok számát, és a ki/bekapcsolhatjuk a zöld színnel ábrázolt strukturális információkat. A találatokat **rendez**hetjük a találati szavak vagy a jobb, illetve bal oldali kontextus szerint. A találatokat **szűr**hetjük, azaz a lekérdezésünk eredményére egy újabb lekérdezést fogalmazhatunk meg a találati szavak melletti adott helyen lévő (megelőző vagy követő) szavakra vonatkozólag.

Több hasznos funkcióval szolgál a **Gyakoriságok** menüpont. Gyakorisági listát készíthetünk szóalakok, illetve a megjelenési évszám szerint. Utóbbi esetben a fejlécre kattintva tudjuk évszám szerinti sorrendbe tenni az eredendően gyakoriság szerint rendezett összesítést. A gyakorisági listán az egyes bejegyzések előtt álló **p** a bejegyzésre vonatkozó pozitív példák (amiben szerepel az adott szó), az **n** pedig a bejegyzésre vonatkozó negatív példák (amiben nem szerepel az adott szó) konkordanciájához visz. A **Gyakoriságok** menüpontra kattintva egy összetett, sokféle gyakorisági lista készítésére szolgáló felületre jutunk. Nagyon hasznos funkció, hogy a találati szavakat megelőző (beállítás: **1L**) vagy követő (**1R**) szavakból is készíthetünk gyakorisági listákat. Ezen kívül **kollokáció**kra is kereshetünk az erre szolgáló külön eszközzel.

A felület angol nyelven is hozzáférhető, a kívánt nyelvet a jobb alsó sarokban választhatjuk ki.

Reguláris kifejezések és a CQL

Az említett funkciókon felül a korpuszkezelő rendszer talán leghasznosabb lehetősége mégis az, hogy használhatjuk az ún. **CQL**-t (Corpus Query Language, korpuszlekérdező nyelv) a korpuszban való kutakodáshoz.

A CQL egy speciális formális lekérdezőnyelv, mely arra szolgál, hogy segítségével – tekintettel a korpusz adatstruktúrájában lévő összes finomságra – részleteiben megfogalmazhassuk a lekérdezésünket, és ezáltal hozzáférjünk a korpuszban lévő információ teljességéhez és minden eleméhez.

Buzdítjuk az olvasót a CQL és a hozzá szükséges ún. reguláris kifejezések megismerésére és használatára, a bemutatott példák kipróbálására. Ebben a fejezetben a cél az, hogy megmutassuk, hogy érdemes ezzel behatóbban foglalkozni. Teljeskörű leírást nem adunk, csak annyit, amennyi a tanulmány példáinak megértéséhez szükséges. A további részletek tekintetében lásd a NoSkE angol nyelvű dokumentációját (Rychlý 2007).

A reguláris kifejezésekről (röviden: regkif) e tanulmány céljaira elég annyit tudni, hogy ezek **bizonyos tulajdonságoknak megfelelő karaktersorozatok** megadására, megragadására szolgálnak. A reguláris kifejezések karakterekből állnak, bennük a karakterek általában „saját magukat" jelentik: az `alma` regkif az *alma* szóra illeszkedik. Bizonyos karaktereknek viszont sajátos jelentésük van: a `.` (pont) speciális karakter például tetszőleges karakterre illeszkedik, azaz a `tejf.l` regkif a *tejföl*, *tejfel*, de még az esetlegesen előforduló *tejfül* stb. alakokat is megragadja. A leggyakoribb speciális karakterek az 1. táblázatban szerepelnek, néhány további példa pedig a 2. táblázatban.

1. táblázat:
A reguláris kifejezésekben használt leggyakoribb speciális karakterek

speciális karakter	jelentés
`.`	tetszőleges (nem szóköz) karakter
*	a megelőző karakterből 0 vagy több
`+`	a megelőző karakterből 1 vagy több
`?`	a megelőző karakter elhagyható
`[ ]`	`[ab]` = *a* vagy *b* karakter

2. táblázat:
Néhány reguláris kifejezés

	regkif	értelmezés
1.	`alma`	*alma*
2.	`tejf.l`	*tejföl*, *tejfel*, *tejfül*, *tejfal* stb.
3.	`.+`	tetszőleges karaktersorozat, bármi, minden szóra illeszkedik
4.	`.+bb`	tetszőleges karaktersorozat a végén 2 db *b*-vel = *bb*-re végződő szó
5.	`nélk[üűű]l`	*nélkül*, *nélkűl* vagy *nélkűl*

Tekintsük a 2. táblázat negyedik példáját, a `.+bb` -t. Fontos érteni a `.+` működését: a `.` tetszőleges (nem szóköz) karaktert jelöl, a + pedig azt jelenti, hogy az **előtte lévő** dologból egy vagy több, azaz a kettő együtt azt jelenti, hogy tetszőleges karakterből egy vagy több, azaz ez a regkif minden szóra illeszkedni fog. A `.+bb` – tetszőleges *bb*-re végződő karaktersorozat – pedig alkalmas arra, hogy megragadjuk a középfokú melléknéveket, ez a regkif jó eséllyel kizárólag középfokú mellékneveket fog eredményezni. A 2. táblázat ötödik példája a *nélkül* szóban lévő *ü* betű MTSz.-ben előforduló három írásváltozatát fedi le.

A CQL egy szinttel továbblép, ha tetszik, tekinthetjük egyfajta (a karakterek helyett) a szavak fölött definiált reguláris kifejezésnek is: **bizonyos tulajdonságoknak megfelelő szósorozatok** megragadására szolgál. A szintaxisa más, néhány speciális karakternek (pl.: `[]`) is más a szerepe. A szögletes zárójelpár itt az egy szóra (vagyis tokenre, ami a szavak és írásjelek szokásos összefoglaló elnevezése) vonatkozó megkötéseket foglalja egybe. A szavaknak lehetnek attribútumai (szótő, morfológiai elemzés stb.), ezekre vonatkoznak a megkötések, melyeket a 3. táblázatban látható módon adhatunk meg. Elemzetlen korpusz lévén az MTSz.-ben csak egy (*word* elnevezésű) attribútum van, maga a szóalak. A 4. táblázatban CQL kifejezésekre láthatunk néhány példát.

3. táblázat:
A CQL legfontosabb elemei

forma	jelentés
`[..]`	egy tokenre vonatkozó megkötések
`x="y"`	megkötés: `x` attribútum értéke legyen `y`
`x!="y"`	megkötés: `x` attribútum értéke *ne* legyen `y`
`&`	*és* kapcsolat megkötések között

4. táblázat:
Néhány CQL kifejezés

	CQL kifejezés	értelmezés
1.	`[] []`	két tetszőleges egymást követő szó
2.	`[word="majd"]`	a *majd* szó
3.	`"majd"`	a *majd* szó (egyszerűsített megadással)
4.	`[word!="a.+"]`	nem *a*-val kezdődő szó

A 4. táblázat első példájában látható üres szögletes zárójelpár azt jelenti, hogy nincs megkötés az adott szóra, azaz tetszőleges szóra illeszkedik. Az egymást követő szavakra vonatkozó zárójelpárokat – ahogy reguláris kifejezések esetén a karaktereket – egymás után írva építhető fel egy CQL kifejezés. Az első példában megfogalmazott lekérdezésnek tehát tetszőleges egymást követő két szó megfelel. A második példa annyit mond, hogy a szóalak – amint említettük, ez az MTSz.-ben a *word* attribútum –, legyen *majd,* azaz itt egyszerűen a *majd* szóra keresünk rá. A harmadik példa egy kényelmi lehetőséget mutat: a felületen a *word* attribútumot alapértelmezettnek beállítva megspórolhatjuk a szögletes zárójelpár és az attribútumnév kiírását. Ez tehát ugyanazt eredményezi, mint a második példa. A negyedik példa mutatja be, hogy egy CQL kifejezés belsejében megjelennek a sima reguláris kifejezések, mégpedig a megkötések jobb oldalán az idézőjeleken belül. Az `a.+` regkif jelentése '*a*-val kezdődő szó', ezt a tagadást kifejező `!=` változtatja az ellenkezőjére.

Az alábbi példában láthatjuk, hogy a karakterszintű regkif operátorok hogyan használhatók a CQL-ben a szavak szintjén:

```
[word="nem"] [word="kellett"] [word="volna"]? [word=".+ni"]
```

Itt egy olyan szósorozatot keresünk, ami úgy kezdődik, hogy *nem kellett,* ez után a `?` operátornak megfelelően vagy ott van, vagy nincs ott egy *volna*, végül egy *ni*-végű

szó következik, ami ebben a kontextusban persze jó eséllyel főnévi igenév lesz. Egy CQL kifejezésben tehát két szinten jelennek meg a reguláris kifejezések operátorai: karakterekre vonatkoztatva (ahogy fent a +) az idézőjelek belsejében, és tokenekre vonatkoztatva (ahogy fent a ?) a záró szögletes zárójelhez tapadva. Ennek megértéséhez hasznos tisztában lenni a reguláris kifejezések működésével.

Lekérdezés és vizsgálódás az MTSz.-ben

Már a korábbi példáknál is láttuk, hogy elemzetlen korpuszban hogyan tudunk bizonyos (morfológiai) tulajdonságokkal bíró szavakat CQL segítségével megragadni. Ez elemzett korpusznál is nagyon hasznos, de elemzetlennél elengedhetetlen. Azt fogjuk látni, hogy a CQL segít hozzá ahhoz, hogy elemzetlen korpuszban is megtaláljuk az adott kutatási kérdéshez a releváns találatokat.

Az MTSz.-keresőfelület használatának illusztrálására most tekintsünk át egy konkrét keresést. Ennek során alkalmazni fogjuk az eddig elmondottakat: a reguláris kifejezéseket, a CQL-t és a NoSkE korpuszkezelő számos funkcióját.

Feladat: keressünk olyan mintázatot (elemzetlen korpuszban!), melyben **tárgyesetű szót követ múlt idejű E/3 ige.** A kívánt szavak legegyszerűbb formai tulajdonságaiból kiindulva megfogalmazhatunk egy lekérdezést, ami alkalmas az ilyenfajta szókapcsolatok megragadására:

```
".+t" ".+tt"
```

A lekérdezés szerint olyan szópárokat keresünk, ahol *t*-végű szót *tt*-végű szó követ. Megjegyezzük, hogy ez a nagyon egyszerű lekérdezés természetesen nem képes az összes kívánt szókapcsolatot megragadni, valamint nem kívánt szókapcsolatok is meg fognak jelenni az eredményben. Az első probléma kezelésére el lehet gondolkodni a múlt idejű E/3. igék felszíni tulajdonságain, és esetleg további mássalhangzókat is meg lehet engedni a szó utolsó előtti pozíciójában (pl.: `".+[djlnrt]t"`). A második probléma megoldására megpróbálkozhatunk bizonyos szavak kizárásával (pl.: `".+t"` helyett `[word=".+t" & word!="most"]`), de ez a probléma kevésbé fontos, mert a találatok közül utólag még kiszűrhetjük a nem megfelelőket. Ez egy általában fontos elv: ha korpuszban keresünk, akkor próbáljuk meg úgy megadni a lekérdezést, hogy lehetőleg minden kívánt nyelvi adatot lefedjen akár azon az áron is, hogy jelentős mennyiségű irrelevéns találatunk lesz, mert az eredményt utólag még szűrhetjük, de a lekérdezéskor elveszített példákat utólag már nem tudjuk pótolni.

A fenti lekérdezéssel a 30 millió szavas MTSz.-ből több mint 86 000 találatot kapunk. Ez bőséges adatmennyiség, amit kézzel átnézni már nem feltétlenül lehetséges. Ilyenkor bizonyulnak hasznosnak a NoSkE korpuszlekérdező különféle funkciói. Az adatok egy aspektusának feltárására például el lehet végezni a felületen következő lépéseket:

1. Futtassuk le a fenti `".+t" ".+tt"` lekérdezést az MTSz. felületén: http://clara.nytud.hu/mtsz. A lekérdezés típusa CQL legyen.
2. Készítsünk gyakorisági listát a szóalakokból (menü: **Gyakoriságok / szóalakok**). A listán számos kívánt kifejezés megjelenik: *részt vett*, *szerepet*

játszott, *pillantást vetett*, *annyit mondott*, *erőt vett*, *mozdulatot tett*, *kezet fogott* stb.

3. Válasszuk ki az *erőt vett* kifejezést, vizsgáljuk tovább ezt. A sor elején álló **p**-re kattintva megkapjuk azt a 122 találatot, melyben ez a kifejezés szerepel.
4. Milyen szó áll rendszerint ez után a kifejezés után? Készítsünk gyakorisági listát a kifejezést követő szavakból (**Gyakoriságok** majd **1R**). Amint várjuk, a leggyakoribb a *rajta* és a *magán* lesz.
5. Válasszuk ki az *erőt vett rajta* kifejezést a *rajta* előtt álló **p**-re kattintva.
6. Milyen dolgok vesznek erőt az emberen? Rendezzük a találatokat a jobb oldali kontextus szerint (**Rendezés / jobb**).

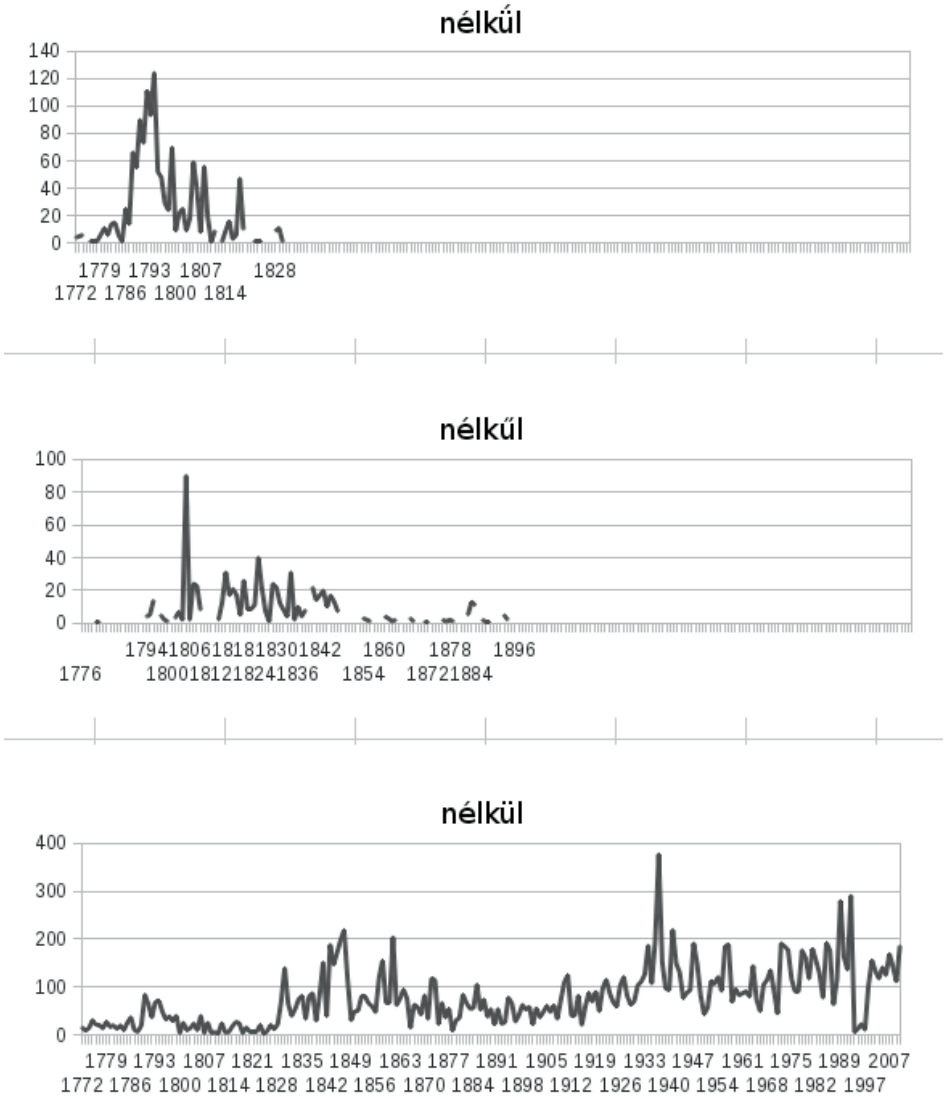
A korpuszvizsgálat eredményeként a következő szavakat kapjuk: *félelem, féltékenység, habozás, kacagás, kishitűség, kíváncsiság* stb. Ez egy szemantikailag többékevésbé koherens (Sass 2011: 51) szócsoport: érzések, azon belül is inkább negatív tartalmú érzések, hangulatok.

A fenti műveletsorozat tipikus példa arra, ahogy a korpuszban való kutatást el lehet kezdeni. Látjuk, hogy elemzetlen korpusszal dolgozva is milyen érdekes jelenségeket tudunk feltárni, milyen hasznos felismerésekre juthatunk a megfelelő módszerek alkalmazásával.

Diakrón vizsgálat: a *nélkül* helyesírása

Egy történeti korpuszban fontos funkció a diakrón vizsgálatok elvégzésének lehetősége. A NoSkE funkcióinál említettük, hogy évszám – a nyelvi adat keletkezési ideje – szerinti gyakorisági listát is lehet készíteni a lekérdezések eredményéből. Nézzük meg, hogy hogyan lehet ennek segítségével alapvető diakrón vizsgálatokat végezni 240 év távlatában.

Az itt bemutatott módszer egy kis utómunkálatot igényel. Először évszám szerinti gyakorisági listát készítünk, kimentjük, áttesszük táblázatkezelő programba, és idő szerinti grafikont készítünk az előfordulási számokból (2. ábra).



2. ábra:
A *nélkül*, illetve a benne lévő *ü* helyesírásának változása az MTSz.-ben. A felső két grafikon a régies *nélkűl* és *nélkűl,* az alsó pedig a mai helyesírásnak megfelelő *nélkül* alak évenkénti előfordulási számát mutatja az MTSz.-ben.

A 2. ábrán látszik, hogy a 18. század végén mindhárom – két pont plusz vesszős, hosszú *ű*-s és rövid *ü*-s – alak is előfordult, aztán átvette az egyeduralmat a ma is hasz-

nálatos *nélkül* alak. Az adatok alapján az 1830-as évek elejére tehető az az időpont, amikor a *nélkül* alak gyakorivá válik, és a másik két alak eltűnik. Megállapíthatjuk, hogy az időpont – nem véletlenül – éppen egybeesik az első magyar helyesírási szabályzat, a Magyar helyesirás' és szóragasztás' főbb szabályai 1832-es megjelenésével.

Megjegyzendő, hogy az ábrázolt adatok nincsenek normálva: nem gyakoriságok szerepelnek, hanem előfordulási számok, és ezek persze függnek az adott évre meglévő korpuszanyag nagyságától, mely évről évre változó mennyiségű. A *nélkül* helyesírását érintő említett változás így is jól látszik, ez a probléma inkább a grafikonon látható különböző csúcsok és völgyek magyarázatául szolgálhat.

A most bemutatott nagyon egyszerű, nem is nyelvi, hanem helyesírási kérdésnek nincs nagy jelentősége. Nem magára a kérdés tartalmára, hanem a vizsgálati módszerre szeretnénk felhívni a figyelmet. A lényeg az, hogy hasonló módon lehet elindulni bármilyen kérdés diakrón korpuszalapú vizsgálatával.

Korpuszalapú gondolkodás: *pimasz*

A korpuszok a nyelvi adatok forrásaként arra szolgálnak, hogy segítségükkel nyelvészeti kérdésfelvetéseket, hipotéziseket alátámasztani vagy cáfolni lehessen. Ha szembetalálkozunk egy nyelvészeti állítással, akkor ha rendelkezésre áll a megfelelő korpusz, azonnal ellenőrizhetjük az állítás igazságtartalmát, megfelelőségét. Kialakítható egy olyan hozzáállás, gondolkodásmód – nevezzük **korpuszalapú gondolkodás**nak –, hogy amikor felmerül egy ilyen állítás vagy kérdés, akkor készségszinten, természetes módon nyúljunk a korpuszhoz, és ott keressünk választ. Ezt szeretnénk még egy – szintén az időbeliséget érintő – esettanulmányban bemutatni.

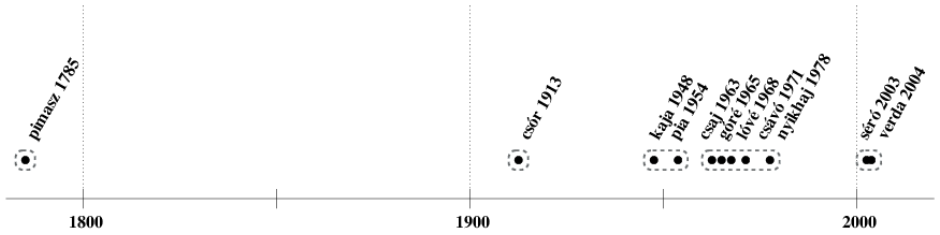
A nyelvtörténeti kutatások újabb eredményei IX. című konferencián meghallgathattam egy cigány jövevényszavakkal kapcsolatos előadást. Az előadásból írásos változat jelen kötetbe nem készült, a vonatkozó kutatásról lásd: Kresztyankó (2016). Egy ponton cigány jövevényszavaknak a következő listáját mutatta be az előadó:

csaj, *csávó*, *csór*, *gádzsó*, *gizda*,
góré, *kaja*, *kéró*, *lóvé*, *nyikhaj*,
pia, *pimasz*, *séró*, *verda*

Nekem első ránézésre úgy tűnt, hogy az egyik szó élesen elüt a többitől, mégpedig a *pimasz*. Az önkéntelen gondolatom az volt, hogy esetleg nem is cigány eredetű ez a szó.

Kakuk Mátyás (1993: 201) megad a *pimasz*-hoz egy lehetséges etimológiát a magyar *arcátlan* felépítéséhez hasonlóan egy cigány fosztóképzőből és az *arc* jelentésű szóból levezetve. Ezt Schirm Anita (2006: 156) ha kétségbe nem is vonja, de nem látja kellően alátámasztottnak, arra hivatkozva, hogy a TESz. a szó eredetét ismeretlenként szótározza. A témához kapcsolódó fontos módszertani kérdéseket tárgyal újabban Arató (2015).

Mit mondhat a kérdésről a korpusz? A módszer jelen esetben ennyi volt: egyszerű kereséssel rákerestünk a megadott szavakra, és megnéztük, hogy mely évből származik az első MTSz.-beli megjelenésük. Az eredmény a 3. ábrán látható.



3. ábra:
Cigány eredetű szavak első megjelenése az MTSz.-ben

Azt látjuk, hogy azon túl, hogy jelentős eltérések vannak az első előfordulás évében, van egy óriási szünet a *pimasz* és az összes többi szó között. Az adatok alapján a szavakat feloszthatjuk az MTSz.-ben való első előfordulásuk szerint csoportokra – lásd a szürke körberajzolásokat az ábrán –, az utolsó csoporthoz *(séró*, *verda)* hozzávehetjük azokat a szavakat, amelyek nem fordulnak elő az MTSz.-ben: *gádzsó*, *gizda* és *kéró*.

A korpuszvizsgálat alátámasztja a hipotézist, hogy a *pimasz* kakukktojás. A helyzet pontos tisztázása persze további vizsgálatokat igényel, de valóban úgy tűnik, hogy ez a szó legalábbis jóval régebbi átvétel (vagy esetleg nem is cigány eredetű). Ezenfelül a *pimasz* – a többivel összehasonlítva – mintha sokkal inkább a magyarba teljesen beépült, köznyelvi, stílusérték nélküli szó lenne. Az is felvethető, hogy a 3. ábrán látható időtengely mentén felállítható egy „köznyelviségi" skála: minél régebbi ezen egy szó, annál inkább köznyelvinek érezzük. A *pimasz* teljesen köznyelvinek hat, a *csór* kicsit kevésbé, a *csaj* még kevésbé, a *gizda* esetében meg talán már elgondolkodunk azon, hogy mit is jelent pontosan. Ez a felvetés aztán további korpuszalapú vizsgálatokkal – képzők hozzáilleszthetősége, közeli szinonimák gyakorisága, kollokációs különbségek – megtámogatható.

Továbbra is hangsúlyozzuk, hogy jelen tanulmány célja nem az adott nyelvészeti probléma teljeskörű körüljárása, hanem az, hogy megmutassuk a korpuszalapú gondolkodás hasznosságát, és bemutassuk, hogy alkalmas eszközökkel mi minden releváns információval szolgálhat a korpusz. Kevésbé a tartalomra, sokkal inkább a módszerre tessék figyelni.

Kitekintés: Nemzeti Korpuszportál

Az MTSz. tagja a Nemzeti Korpuszportál (NKP) kezdeményezésnek (Sass 2016). Ennek az a célja, hogy összegyűjtse a magyar nyelvű, szóalapú online keresővel rendelkező korpuszokat, távlatilag pedig az, hogy minden lekérdezőfunkciót elérhetővé tegyen minden korpuszhoz. Az NKP ez a célt éppen a NoSkE korpuszkezelő általánossá, sztenderddé tételével kívánja megvalósítani: a gazdag funkcionalitású NoSkE alkalmas egységes platformnak látszik mindenféle korpusz számára. Segítségével lényegében tetszőleges korpuszhoz lehet a leírthoz hasonló felületet készíteni.

Jelen tanulmányban ismertetett módszerek ezáltal egyre szélesebb körben alkalmazhatóvá válnak. A reguláris kifejezések és a CQL jelenleg is számos NKP-s korpusz esetén használhatók, a Magyar Nemzeti Szövegtár új változatában (Váradi és Oravecz

2014) pedig – mely már most a NoSkE-re épül – lényegében pontosan úgy kereshetünk, ahogy jelen tanulmányban ismertettük.

Hivatkozások

Arató Mátyás 2016: A romani és beás eredetű szavak alapkérdései és alapproblémái, in Bagyinszki Szilvia – P. Kocsis Réka szerk.: *Anyanyelvünk évszázadai* 2, Budapest, ELTE Magyar Nyelvtörténeti, Szociolingvisztikai, Dialektológiai Tanszék, 71–82.

Csengery Kinga 2006: Az elektronikus korpusz, in Ittzés Nóra, szerk.: *A magyar nyelv nagyszótára 1. Segédletek*, Budapest, MTA Nyelvtudományi Intézet.

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

Kakuk Mátyás 1993: A magyar nyelv cigány jövevényszavaiból, *Magyar Nyelv* 89, 196−204.

Kresztyankó Annamária 2016: Cigány jövevényszavak nyelvtanórán, in Bagyinszki–P. Kocsis, szerk.: *Anyanyelvünk évszázadai 2.*, ELTE BTK, Budapest, 215–23.

Rychlý, Pavel 2007: Manatee/Bonito – A modular corpus manager, in *Proceedings of the 1st Workshop on Recent Advances in Slavonic Natural Language Processing*, Brno, Masaryk University, 65–70. https://nlp.fi.muni.cz/trac/noske

Sass Bálint 2011: *Igei szerkezetek gyakorisági szótára – egy automatikus lexikai kinyerő eljárás és alkalmazása*, PhD dolgozat, Budapest, PPKE ITK.

Sass Bálint 2016: Nyelvészeti szövegkeresők, Nemzeti Korpuszportál, *Magyar Tudomány*, 177/7, 798–808.

Schirm Anita 2006: A magyar nyelv cigány eredetű jövevényszavai, *Nyelvtudomány* 2, 149–63.

Schirm Anita 2014a: A deviszont viszontagságai. http://www.nyest.hu/hirek/a-deviszont-viszontagsagai

Schirm Anita 2014b: A diskurzusjelölők stilisztikai és pragmatikai megközelítése, in Dobi Edit – Domonkosi Ágnes – Pethő József szerk.: *Stílusról, nyelvről – sokszínűen, Szikszainé Nagy Irma 70. születésnapjára*, Debrecen, 294–307.

Schirm Anita 2015: Deviszont vannak még érdekességek a deviszont körül. http://www.nyest.hu/hirek/deviszont-vannak-meg-erdekessegek-a-deviszont-korul

Váradi Tamás és Oravecz Csaba 2014: A Magyar Nemzeti Szövegtár egymilliárd szavas új változata, *Magyar Tudomány*, 175/9, 1054–62.