A Magyar Nemzeti Szövegtár új változatáról

Váradi Tamás

varadi@nytud.mta.hu

MTA Nyelvtudományi Intézet

Nyelvtechnológiai és Alkalmazott Nyelvészeti Osztály

Tartalom

- Előzmény
- Motiváció
- Cél
- Fejlesztés
- Eredmény

Előzmény

Magyar Nemzeti Szövegtár (MNSz)

- 1998 és 2001 között készült
- 90-es évek második felének nyelvhasználatából merített reprezentatív minta
- az első, az akkori gyakorlatban is jelentős méretűnek számító, nyelvileg elemzett, hálózati lekérdező felületen bárki számára szabadon hozzáférhető korpusz
- 187 millió szó, határon túli nyelvváltozatokkal kibővített anyag
- több mint 7000 felhasználó, több tucat tanulmány az MNSz adatai alapján

Motiváció

15 évvel később ...

- nyelvi erőforrásokkal szemben támasztott igények jelentős mértékben változtak
  - adatközpontú módszerek/alkalmazások elterjedése és sikeressége a számítógépes nyelvfeldolgozás területén
    $\longrightarrow$ minél több a nyelvi adat, annál jobbak az eredmények
  - fejlett(ebb) nyelvi elemző eszközök
    $\longrightarrow$ jobb minőségű és részletesebb nyelvi elemzés és annotáció
  - reprezentativitás megőrzése
    $\longrightarrow$ a nyelvhasználat újabb és újabb mintavételezése, a nyelvi változatok széles skálájából
- ... az MNSz mára elavulttá vált.

Cél

Megnövelt ...

- *minőség.* A korpusz anyagának minden feldolgozási és elemzési lépésében új, korszerű számítógépes nyelvészeti technológia felhasználása.
- *méret.* A korpusz anyagának bővítése 1000 millió szóra.
- *lefedettség és reprezentativitás.* Újabb mintavétel a mai magyar nyelvhasználatnak a Szövegtárban eddig is szereplő, valamint további változataiból (*„social media”*).

MNSz2: Korszerű nyelvi erőforrás létrehozása, amely színvonalasan szolgálja ki a magyar nyelvi adatokat felhasználó kutatásokat, és az érdeklődő nagyközönséget is.

Cél

Megnövelt ...

- *minőség.* A korpusz anyagának minden feldolgozási és elemzési lépésében új, korszerű számítógépes nyelvészeti technológia felhasználása.
- *méret.* A korpusz anyagának bővítése 1000 millió szóra.
- *lefedettség és reprezentativitás.* Újabb mintavétel a mai magyar nyelvhasználatnak a Szövegtárban eddig is szereplő, valamint további változataiból (*„social media”*).

MNSz2: Korszerű nyelvi erőforrás létrehozása, amely színvonalasan szolgálja ki a magyar nyelvi adatokat felhasználó kutatásokat, és az érdeklődő nagyközönséget is.

Fejlesztés

Anyaggyűjtés

- szerzői jogi kérdések tisztázása
- elegendő metaadat (interneten elérhető szövegek automatikus letöltése nem feltétlen megfelelő)
- automatikus feldolgozhatóság → pdf, OCR nem használható

Előfeldolgozás, szövegnormalizálás

- szöveges tartalom és alapvető dokumentumszerkezet azonosítása
- (közel) duplikátumok és idegen nyelvű szövegrészek kiszűrése

Fejlesztés

Elemzés és annotáció

- részletes morfoszintaktikai elemzés újratervezett automatikus egyértelműsítő architektúrával
  (morfémákra, összetételekre, szótagszerkezetre vonatkozó információk)
- szabványos XML formátum, IOB belső reprezentáció

IOB formátum


```
az          az          DET       D__D      compound=n;;hyphenated=n;;stem=az::DET
                                    BC        BC        az        az        B
angol       angol       MN.NOM    AS_A      compound=n;;hyphenated=n;;stem=angol::MN;;
                                              morphemes=ZERO::NOM       I
                                    BCCBC     BCCBC     angol     angol     I
nyelvű      nyelvű      MN.NOM    AS_A      compound=n;;hyphenated=n;;stem=nyelv::FN;;
                                              morphemes=ZERO::NOM ű::_UKEP
                                    CNCCF     CNCCF     Nelvű     Nelvű     I
szöveg      szöveg      FN.NOM    NS3NN     compound=n;;hyphenated=n;;stem=szöveg::FN;;
                                              morphemes=ZERO::NOM
                                    CFCNC     CFCNC     Söveg     Söveg     I
az          az          DET       D__D      compound=n;;hyphenated=n;;stem=az::DET
                                    BC        BC        az        az        O
irányadó    irányadó    MN.NOM    AS_A      compound=y;;hyphenated=n;;stem=ad::IGE
                                              irány::FN;;morphemes=ZERO::NOM ó::_OKEP
                                    NCBCBCB   NCBCBCB   iráNadó   iráNadó   O
.           .           SPUNCT    __SPUNCT__    __NA__    __NA__    __NA__    __NA__    __NA__
```

Fejlesztés

Elemzés és annotáció

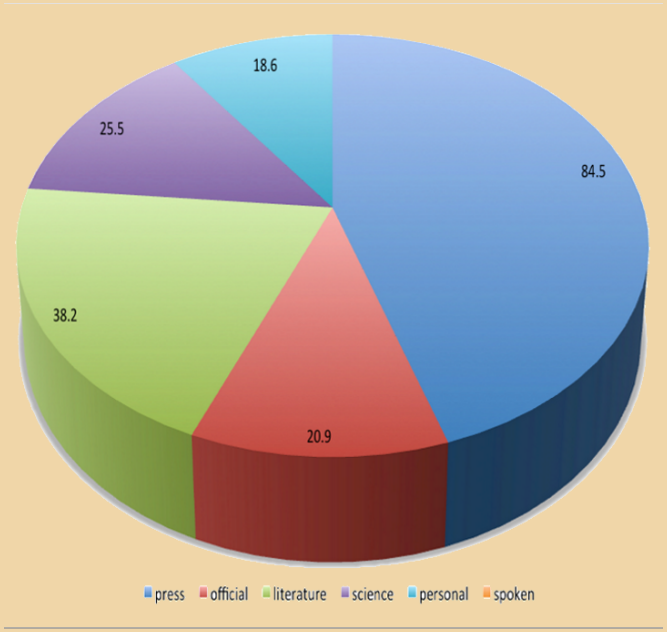
- részletes morfoszintaktikai elemzés újratervezett automatikus egyértelműsítő architektúrával
  (morfémákra, összetételekre, szótagszerkezetre vonatkozó információk)
- szabványos XML formátum, IOB belső reprezentáció

Korpuszkezelő

- robusztus, több milliárd szavas adatbázist kezelni képes
- gyors válaszidő

Eredmény

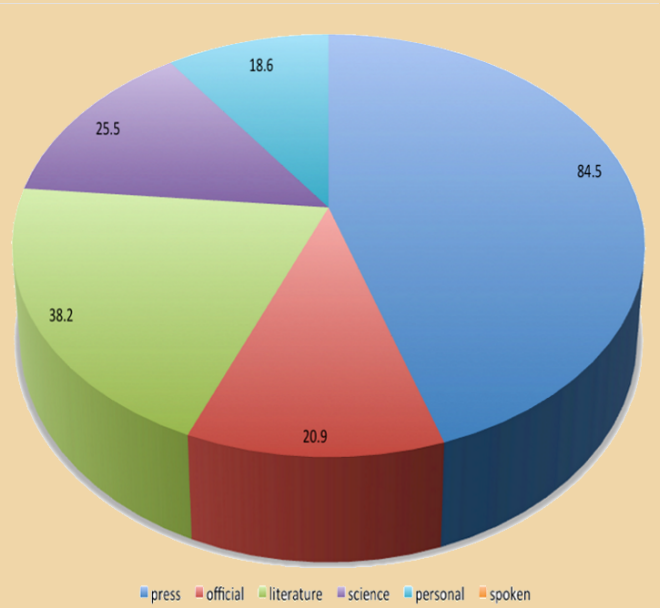
MNSz: 187 m.



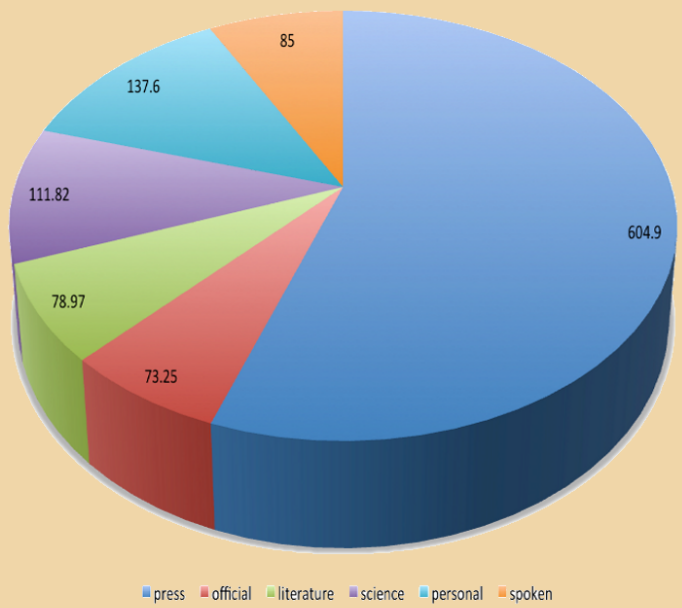
MNSz2 (+MNSz): 1091 m.

Eredmény

MNSz: 187 m.



MNSz2 (+MNSz): 1091 m.



Eredmény

„Intelligens” korpusz

- összetett menüvezérelt keresés a kódolt információ minden részletére
  - morfo(fono)lógiai jelenségek
  - többszavas kifejezések: kollokációk, igei argumentumok
- megjelenítési beállítások: szövegkörnyezet, metaadatok
- megoszlásvizsgálatok, beépített utófeldolgozás (több szintű gyakorisági listák, megelőző eredmények további szűrése és feldolgozása)

"piros ..."

Concordance
Word List

Save
View concordance
Sample
Filter
Frequency
Node tags
Node forms
Doc IDs
ConcDesc

Collocation candidates

Page 1 Go Next >

		Freq	T-score	MI	logDice
p/n	lámpa	292	17.066	9.624	8.804
p/n	kockás	122	11.039	10.752	7.920
p/n	színű	132	11.469	9.170	7.823
p/n	betűs	101	10.046	11.182	7.687
p/n	kék	186	13.566	7.559	7.596
p/n	rózsa	124	11.098	8.217	7.518
p/n	lap	325	17.864	6.779	7.383
p/n	süt	111	10.496	8.040	7.352
p/n	zászló	109	10.394	7.822	7.264
p/n	sárga	126	11.155	7.328	7.194
p/n	folt	85	9.190	8.298	7.138
p/n	alma	73	8.524	8.733	7.041
p/n	zöld	148	12.049	6.704	6.973
p/n	elefánt	63	7.925	9.304	6.925
p/n	szín	123	10.978	6.626	6.817
p/n	ceruza	58	7.601	9.017	6.789
p/n	sapka	60	7.723	8.393	6.749
p/n	arcú	58	7.597	8.680	6.749
p/n	jelzés	71	8.371	7.249	6.661

"piros lámpa"

történt volna . Így eshetett meg , hogy **piros** *lámpánál* az út közepén futottam , s éppen
autótolvajok új stratégiát követnek . Így például a **piros** *lámpánál* hátulról belehajtanak a kiszemelt
Módszerük az , hogy az autópályáról levezető **piros** *lámpánál* megálló gépkocsi mellé berobogó
Felelőtlenség a gyerekeknek rossz példát mutatni , **piros** *lámpánál* átszaladni , a szabályokat nem
 Mindennapi tapasztalatot elevenít fel a **piros** *lámpánál* megrekedt autósról szóló jelenet
össze a motoros rendőrökkel . Ha még a **piros** *lámpán* is áthajtott , és a kerékpár felszereltsége
vezető minden alkalommal megjelent , hatalmas **piros** elemes *lámpát* tartott a kezében , és elmondta
arcom . De az biztos , ha a stúdióban ég a **piros** *lámpa* , akkor mindenki egy kicsit összekapja
és kíséretét szállító konvoj minden egyes **piros** *lámpánál* megállt Lahorban - a szemtanúk
gépkocsioszlop szabályosan megállt minden egyes **piros** *lámpánál* , nem kis megdöbbenést okozva
beavatkozásokat eszközölni , ők csupán várakoznak a **piros** *lámpánál* és adományokat gyűjtenek az autósoktól
" technika . Az előző módszernél a **piros** *lámpánál* álló vezetőt kikényszerítik autójából
például annak , hogy amikor Pesten járt , és a **piros** *lámpánál* egy autóbusz utolérte , a sofőr
kikapcsolása . (A három futó állapotát egy sor **piros** *lámpa* is jelzi . Csakhogy Szaloniki felett

Felhasználói felület

http://mnsz2.nytud.hu

Köszönöm a figyelmet!