



NYELV- ÉS BESZÉDTECHNOLÓGIAI PLATFORM

Magyar Nemzeti Szövegtár



NYELVTUDOMÁNYI INTÉZET
MAGYAR TUDOMÁNYOS AKADÉMIA

<http://corpus.nytud.hu/mnsz>

Szövegtár, számítógépes korpusz

- Írott, vagy lejegyzett beszélt nyelvi adatok elektronikus formában tárolt gyűjteménye.
- Nem archívum, nem feltétlenül egész szövegeket tartalmaz.
- Tartalmaz viszont bibliográfiai adatokat, a nyelvi elemeket jellemző elemzést, jelöli a szerkezeti egységeket (bekezdés, mondat).
- Számítógéppel hatékonyan kezelhető szabványos formátumban, adatbázisként használható.
- Elméleti kutatások és számítógépes nyelvészeti eljárások, alkalmazások nélkülözhetetlen nyersanyaga.
- Értékes információt hordoz az adott nyelvhez kötődő kultúra kutatóinak, társadalomtudósainak számára is.

Hogyan készült?

- Elektronikus források: sajtószövegek, Digitális Irodalmi Akadémia, Magyar Elektronikus Könyvtár, törvények, parlamenti napló, internetes fórumok.
- Előfeldolgozás: bekezdések, mondatok azonosítása, egy-egy szabványos formátum kialakítása.
- Nyelvtani elemzés: elemekre bontás, morfológiai elemzés.
- Egyértelműsítés: többértelműségek feloldása (pl. *járatok.Ige/Fn*)
- Összefűzés, indexálás: lekérdezhető adatbázis kialakítása.
- Lekérdező szoftver: felhasználói felület kifejlesztése.

Felhasználók

Több mint 5000 regisztrált felhasználó:

- Nyelvészek, szótárszerkesztők
- Számítógépes nyelvészek
- Társadalomtudósok
- Nyelv iránt érdeklődők

Intelligens keresés

- Szótó és szóalak keresése
- Keresés szófaj és alaktani jellemzők alapján
- Két szó együttes előfordulása
 - állandósult szókapcsolatok
 - igei vonzatok
- Megjelenítési beállítások
 - kontextus
 - alkorpuszra korlátozás
- Megoszlásvizsgálat

A Magyar Nemzeti Szövegtár

- A mai magyar írott köznyelv általános célú reprezentatív korpusza.
- Számszerűsíthető képet ad a magyar nyelvhasználatról.
- **Intelligens korpusz:**
 - minden szó mellett feltüntet a szótövet, a szófajt és a szó morfológiai elemzését,
 - a lekérdezés nyelvtani jellemző szerint is lehetséges.
- Hálózati lekérdező felületen bárki számára szabadon hozzáférhető.

Az MNSZ összetétele

Mitől *nemzeti*?

Mérete szerint:

- hasonló vállalkozásokkal összemérhető (pl. British National Corpus)

Tartalma szerint:

- nem csak az „írastudó elit” nyelvezete
- nem csak budapesti nyelvhasználat
- határon túli nyelvhasználat is (*Kárpát-medencei Magyar Korpusz*)

	magyarországi	szlovákiai	kárpátaljai	erdélyi	vajdasági	Összesen (m. szó)
sajtó	71,0	5,7	0,7	5,5	1,5	84,5
szépirodalom	35,5	1,4	0,4	0,8	0,2	38,2
tudományos	20,5	2,3	0,7	1,6	0,3	25,5
hivatalos	19,9	0,2	0,3	0,6	0,1	20,9
személyes	17,8	—	0,4	0,4	0,1	18,6
összesen	164,7	9,5	2,5	8,9	2,0	187,6

1. táblázat. Nyelvváltozatok és stílusrétegek számszerűen.

Az MNSZ belülről

```
<h.title>Sinistra körzet</h.title>
<h.author>Bodor Ádám</h.author>
```

```
...
<s>
<w lemma="most" msd="Adv" ctag="R">Most</w>
<w lemma="azonban" msd="Con" ctag="C">azonban</w>
<w lemma="járatlan" msd="A.NOM" ctag="AS_A">járatlan</w>
<w lemma="ösvény" msd="N.PL.SUP" ctag="NP3NP">ösvényeken</w>
<c lemma="," msd="WPUNCT" ctag="WPUNCT">,</c>
<w lemma="a" msd="Det" ctag="D">a</w>
<w lemma="hegyi" msd="A.NOM" ctag="AS_A">hegyi</w>
<w lemma="vadász" msd="N.PL.NOM" ctag="NP3NN">vadászok</w>
<w lemma="útjelzés" msd="N.PSe3i.ACC" ctag="NP3NAS3">útjelzéseit</w>
<w lemma="követ" msd="V.HIN" ctag="R_V">követve</w>
<w lemma="egyenesen" msd="Adv" ctag="R">egyenesen</w>
<w lemma="titkos" msd="A.NOM" ctag="AS_A">titkos</w>
<w lemma="kilátóhely" msd="N.PSe3.NOM" ctag="NS3NNS3">kilátóhelye</w>
<w lemma="felé" msd="NU" ctag="RP">felé</w>
<w lemma="tart" msd="V.Me3" ctag="VS3PI">tartott</w>
<c lemma="." msd="SPUNCT" ctag="SPUNCT">.</c>
</s>
```

A screenshot of the MNSZ search interface. The search results show a list of documents related to 'anyországi támogatás'. The interface includes search filters, a list of search results with details like 'Régió: erdélyi', 'Számszerűség: sajtó', and 'Cím: Kárpátalja: A legjelentősebb vívmány?'. Below the search results, there are two pie charts showing the distribution of documents by region and by document type.

A találatok stílusrétegek szerinti megoszlása:

sajtó	38 db	0,45 db / millió szó
szépirodalom	0 db	0,00 db / millió szó
tudományos	9 db	0,35 db / millió szó
hivatalos	3 db	0,14 db / millió szó
személyes	0 db	0,00 db / millió szó

A találatok régiók szerinti megoszlása:

magyarországi	22 db	0,13 db / millió szó
szlovákiai	4 db	0,42 db / millió szó
kárpátaljai	2 db	0,80 db / millió szó
erdélyi	18 db	2,03 db / millió szó
vajdasági	4 db	1,96 db / millió szó