

Korpuszokról adatgazdászoknak

a Magyar Nemzeti Szövegtár és azon túl

adatgazdász találkozó
2026. január 23.

Sass Bálint

ELTE Nyelvtudományi Kutatóközpont
Lexikológiai Intézet
`sass.balint@nytud.hu`

korpusz



korpusz = sok szöveg :)

van hozzá lekérdező, aminek segítségével szavakat keresünk benne
számítógépes rendszer

nyelvészeti releváns kérdéseket szeretnénk feltenni – *annotáció*

(adott tulajdonságú szavak sorozatai, mi? milyen gyakran? milyen környezetben?)

korpusz

- korpusz = *tokenek* sora
- *vertikális* formátum
- token: alapegység → alapvetően: szó *vagy* írásjel

először, _hogy (← ez nem 2 egység, hanem 3)

- kiegészítő adatok: *annotáció* – ez tartalmazza a nyelvi tudást

→ egy *táblázat*:

- * egy token = egy sor
 - * tokenszintű annotáció oszlopokban: *attribútumok*
 - * struktúrák XML-szerűen
- ez a NoSketchEngine (NoSkE)
korpuszkezelő rendszer alapformátuma

elemzetlen korpusz

szóalak

jelent
meg
először

,
hogy
Francis
Crick
és

elemzetlen korpusz = **tokenek sora**

elemzett korpusz

szóalak	szó/íj	szófaj	szótő	NE
jelent	w	ige	jelenik	○
meg	w	ik	meg	○
először	w	hsz	először	○
,	p	p	,	○
hogya	w	ksz	hogya	○
<i>Francis</i>	w	fn	Francis	B-PERS
<i>Crick</i>	w	fn	Crick	I-PERS
és	w	ksz	és	○

elemzett korpusz = **tokenek + annotáció** ~ táblázat

elemzett korpusz

szóalak	szó/íj	szófaj	szótő	NE
---------	--------	--------	-------	----

<s>

← **struktúra**

...

jelent	w	ige	jelenik	○
meg	w	ik	meg	○
először	w	hsz	először	○
,	p	p	,	○
hogy	w	ksz	hogy	○
<i>Francis</i>	w	fn	Francis	B-PERS
<i>Crick</i>	w	fn	Crick	I-PERS
és	w	ksz	és	○

</s>

elemzett korpusz = **tokenek + annotáció** ~ táblázat

elemzett korpusz

szóalak	szó/íj	szófaj	szótő	NE
---------	--------	--------	-------	----

<s id="5">

← **struktúra-annotáció**, *metaadat*

...

jelent	w	ige	jelenik	O
--------	----------	------------	----------------	----------

← **token-annotáció**

meg	w	ik	meg	O
-----	---	----	-----	---

először	w	hsz	először	O
---------	---	-----	---------	---

,	p	p	,	O
---	---	---	---	---

hogy	w	ksz	hogy	O
------	---	-----	------	---

Francis	w	fn	Francis	B-PERS
---------	---	----	---------	--------

Crick	w	fn	Crick	I-PERS
-------	---	----	-------	--------

és	w	ksz	és	O
----	---	-----	----	---

</s>

elemzett korpusz = **tokenek + annotáció** ~ táblázat

adatifajták korpuszokban

- **szöveg-adat** – tokenek
- **annotáció-adat** – token-annotáció, struktúrák, struktúra-annotáció

span: hány tokenre terjed ki az annotáció

alapvetően:

$\text{span} = 1 \rightarrow$ token-annotáció

$\text{span} > 1 \rightarrow$ struktúra-annotáció

de:

$\text{span} > 1$ token-annotációval: IOB-kódolás

- **metaadat** – amikor a *span* a teljes dokumentum/korpusz vagy egy mondat struktúra-annotációját is hívhatjuk metaadatnak?

?

méretek

szószám	könyv	tárhely	példa
10000	5mm	0,1 – 1M	KorKor
100000	5cm	1 – 10M	BUSZI, MoMa, KED
1 millió	50cm	10 – 100M	Ómagyar
10 millió	5m	0,1 – 1G	Mtsz, MMKK
100 millió	50m	1 – 10G	MNSZ1 (2005)
1 milliárd	500m	10 – 100G	MNSZ2 (2014)
10 milliárd	5km	0,1 – 1T	MNSZ3 – készül
100 milliárd	50km	1 – 10T	AI-tanítás (nyers szöveg)

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

Magyar Nemzeti Szövegtár 2. verzió

1 milliárd szó

„a magyar nyelv reprezentatív korpusza”: szépirodalom, hivatalos, tudományos, sajtó, személyes, rádiós; hazai és határon túli anyag
részletes annotáció

hozzáférés: online regisztráció

demó:

- *ember* mint szótő
- *körüliültük*
- [lemma_cv="CBCCNC"]
- közös, jó, szakmai, tudományos, alkalmi → minek a közös jelzői?

Mazsola – ugyanaz az adat, más típusú lekérdező
igék és bővítményeik vizsgálatára

reprezentáció:

A lány vállat vont. → ige=von alany=lány tárgy=váll

demó:

- *eszik -t*
- *hagy -t*

Budapesti Szociolingvisztikai Interjú

270000 szó

részletesen lejegyzett *beszélt nyelvi* korpusz
más, gazdag annotáció, normalizálás

hozzáférés: offline regisztráció aláírással

demó: *elképzeltem*

bizonyos dógokban □ *mmm tát, hogy ööö*

*lustább annál, mint amilyennek elkép*zel*tem, ...*

nyelvészeti szempontból nagyon is érdekes

adatgazdász szempontból valszeg csak más adattartalom

KorKor

4/6

anafora- és koreferencia-annotáció

25000 szó

ellenőrzött kézi annotálás

hozzáférés: szabadon letölthető

demó: `coref/test/28`

kézi annotálás → nagyon értékes adat!

`https://github.com/vadno/korkor\_pilot`

11/20

Moldvai magyar korpusz

200000 szó

a korpusz tartalma hanganyagként is hozzáférhető

hozzáférés: nyílt lekérdező + szabadon letölthető

demó: *ura* (33-34. találat)

**ez a hozzá tartozó hanganyag miatt
talán adatgazdász szempontból is érdekes**

Magyar Mozgókép Korpusz

11 millió szó

filmfeliratok időpont-annotációval

cél magyar mint idegen nyelv oktatásához élőnyelvi példák filmekből
plusz a magyar kultúra bemutatása

hozzáférés: nyílt lekérdező

demó: *mit tudom én* – a Magyar vándor c. filmben

**magukat a filmeket nem birtokoljuk,
de használatba tudjuk őket venni hivatkozással**

gold-standard vs silver-standard adat

gold-standard adat: ellenőrzött, kézi annotálás, profi annotátorok, világos útmutató szerint, inter-annotátor agreement mérése és közlése

KorKor

BUSZI és MoMa esetén a lejegyzés

silver-standard adat: „majdnem” kézi minőségű

pl. 97,5%-os pontosságú automatikus annotáció

pl. wikipedia link, ami nagybetűs → tulajdonnév

elemzések

eredeti adat vs származtatott adat

- **eredeti** adat = nem generálható automatikusan

amit csak adott forrásból (adatközlőtől)

és/vagy adott időpontban lehet(ett) rögzíteni

és/vagy nagyon nehéz begyűjteni

→ egyedi érték, gondosan tárolandó

- **származtatott** adat = automatikusan generálható

alapvetően egy „fölslegesen” tárolt, redundáns adat

elvileg elég lenne a módszert/scriptet tárolni, ami előállítja

→ elvileg nem szükséges tárolni

eredeti adat vs származtatott adat

mégis: a származtatott adat is lehet tárolva

végülis az MNSZ2 teljes elemzése generálható,
de nyilván tárolni fogjuk, mert így jó a kereséshez

kulcs: *mennyi erőforrás felhasználásával állítható elő
meglévő adatból a származtatott adat* – ha sok, inkább eltároljuk
ha generált adatot tárolunk, a *karbantartás* plusz teher:
adatváltozáskor újra kell generálni

az igény szerint (**on demand**) generálódó adat – nem tárolódik
gondolatom: annak is a gazdái vagyunk
dinamikus honlap, scrabble-projekt, szerkezetár
ami kihozható, az „benne van” – pl. *asztalt*

a korpusz kutatási adat?

egy adatgazdászok előadásán hallottam ...

kutatási adat:

ami kísérlet, kutatás folyamán jön létre

ez alapján tehát a korpusz nem kutatási adat?

akkor mi? :)



hozzáférhetőség

elv: **azért vannak a korpuszok, hogy használják őket**
mindent jó lenne közzétenni a lehető legszabadabb módon

1. szabadon letölthető adat

KorKor, MoMa

2. nyílt lekérdező

MoMa, MMKK

3. zárt lekérdező, online regisztráció automatikusan ~ 50% veszteség

MNSZ2

4. zárt lekérdező, offline regisztráció aláírással ~ 70% veszteség

BUSZI

minél zártabb, annál inkább elvész a kutatás számára

hozzáférhetőség

időről-időre belefutunk abba, hogy
évtizedekkel korábbi nyilatkozatok, szerződések
korlátozzák a közzétételt

pl. BUSZI, MNSZ2, BEA

→ lehet-e ezen valami módon könnyíteni?



változnak az idők, ma már nem úgy írnánk meg a nyilatkozatokat

ha anonimizálom, akkor közzétehetem?

összefoglalás

- mi a korpusz: tokenek sora + annotáció
- adatfajták, méretek
- MNSZ2, Mazsola, BUSZI, KorKor, MoMa, MMKK
- gold-standard *vs* silver-standard adat
- eredeti adat *vs* származtatott adat
- a korpusz kutatási adat?
- hozzáférhetőség

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

`sass.balint@nytud.hu`