



Bevezetés a nyelvtechnológiába

10.

Korpuszok

(és még néhány dolog, ami eddig kimaradt...)

A korpusznyelvészet alapgondolata

- ☐ Korpusz: „meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi”
- ☐ John R. Firth (1890-1960): „*You shall know a word by the company it keeps*”
- ☐ John M. Sinclair (1933-2007): „*Language cannot be invented; it can only be captured*”
- ☐ Korpusz-alapú módszer: a szövegtörzs segédeszköz, ami empirikus adataival támogatja az intuíciót, mérhetővé teszi a nyelvi jelenségeket, meglévő elméleteket bizonyít/cáfol
- ☐ Korpuszvezérelt módszer: a korpusz maga szolgáltatja az „elméletet”, a nyelvész minden előzetes feltevés és elvárás nélkül fordul az adatokhoz, és minden következtetést kizárólag a korpusz megfigyeléseiből von le
- ☐ Serendipity Principle: jelentős dolog véletlen felfedezése, tipikusan akkor, amikor valami egész másra figyelünk

Régi dilemma: nyelvtankönyv vs. szöveg

Riedl Frigyes: *Simonyi kis nyelvtana* (1882)

„Simonyi új grammatikai módszert akar behozni, könyve inductive halad, azaz a példákból kiindulva tanítja a szabályt, nem pedig dogmatica.

A grammaticát tehát valami olvasmány alapján akarja előadni, úgy hogy a szabályokat a tanár tanítványai közreműködésével vonhatja le ésszerű következtetések útján. Ilyenképp tehát ezen módszer véget vet a lelketlen magolásnak, és azt észfejlesztő inductióval pótolja. Eszerint a szabályok is mélyebben vésődnek be a gyermek emlékezetébe, mert amit magunk találunk, azt jobban tudjuk, mint amit más mond vagy más tanultat velünk.”

A korpusz méretének mérőszámai

- ☐ Hány „token” van benne? = Mekkora a korpusz?
- ☐ Mi egy token?
 - ☐ \$22.50
 - ☐ George W. Bush / George Bush / Bush
- ☐ Korpusznormalizálás
 - ☐ The / the / THE
 - ☐ Calif. / California
 - ☐ MTA / Magyar Tudományos Akadémia
- ☐ Hány „type” van benne?
= Hány különböző szó van benne?
- ☐ A „type”-ok tényleges gyakorisága

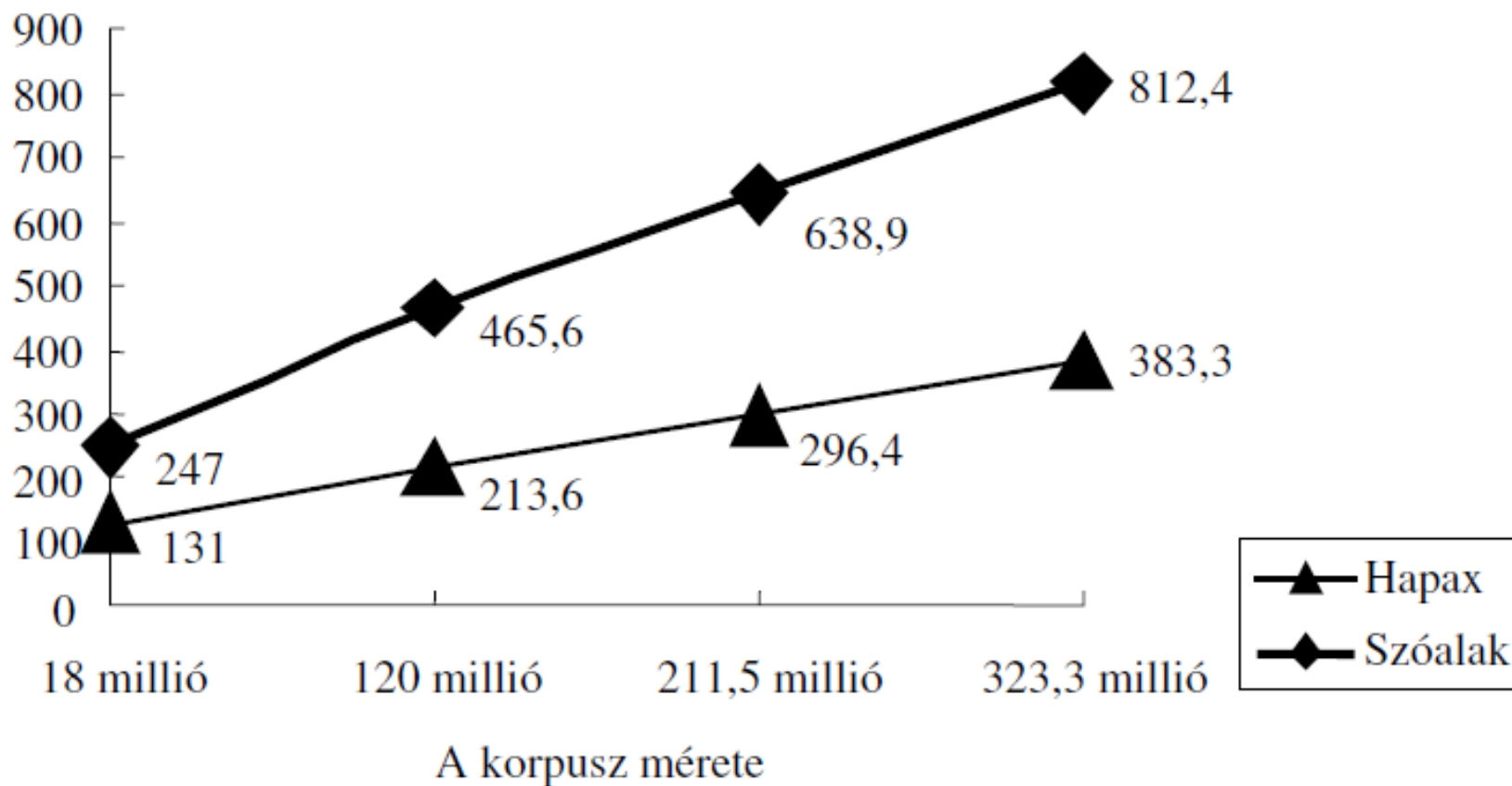
A korpuszok méretéről

(Szirmai M. alapján)

☐ A4 oldal (kettes sorközzel)	250 szó
☐ Gárdonyi: Egri csillagok	135 000 szó
☐ Brown Corpus (1961)	1 000 000 szó
☐ COBUILD Corpus (1987)	18 000 000 szó
☐ British National Corpus (2002)	100 000 000 szó
☐ Magyar Nemzeti Szövegtár (2005)	187 600 000 szó
☐ COBUILD Corpus (1996)	323 000 000 szó
☐ Corpus of Cont. Am. Engl. (2011)	425 000 000 szó
☐ MNSZ 2 (2014)	500 000 000 szó
☐ COBUILD Corpus (2012)	650 000 000 szó
☐ Pázmány Korpusz (2016)	1 200 000 000 szó

A korpuszok méretéről (2)

A szóalakok és hapaxok növekedése a Cobuild Korpuszban



Korpuszannotációk

- ☐ Egyszerű szöveg
- ☐ Különböző annotációk mindenféle hipotézisek teszteléséhez
- ☐ Szófaji egyértelműsítés
- ☐ Név kifejezések kezelése
- ☐ Szintaktikai szerkezetek bejelölése
- ☐ Jelentés-egyértelműsítés
- ☐ Koreferencia-kezelés
- ☐ Dialógus-jegyek



Néhány híres címkézett korpusz

Korpusz	# Tokenek	Megjegyzés
Brown	1 000 000	címkézett, kiegyensúlyozott
Susanne	120 000	a Brown szintaktikailag elemzett részhalmaza
Lancaster (-Oslo-Bergen)	1 000 000	„a UK válasza a Brown-ra”
Penn Treebank	2 000 000	szintaktikailag elemzett
British National Corpus (BNC)	100 000 000	szófajilag egyértelműsített
Szeged Korpusz	1 200 000	kiegyensúlyozott, morfológiailag és (2. verzió) szintaktikailag elemzett
Magyar Nemzeti Szövegtár	187 600 000	szófajilag egyértelműsített
Magyar Nemzeti Szövegtár 2	500 000 000	szófajilag egyértelműsített
Pázmány Korpusz	1 200 000 000	szófajilag egyértelműsített, NP-k

A Brown Corpus

- ❑ Standard American English
- ❑ Az első modern korpusz
(Francis and Kucera, 1961)
- ❑ 500 szöveg, darabja 2000 szó hosszú
- ❑ Amerikai könyvek, újságok, folyóiratok
- ❑ 15 témakör: tudományos-fantasztikus, regény, sajtó, tudományos művek
- ❑ POS-taggelt: 87 osztály



A CLAWS címkekészlet

(Constituent Likelihood Automatic Word-tagging System)

ABL pre-qualifier (quite, rather, such...)

ABN pre-quantifier (all, half)

...

AT singular article (a, an, every)

ATI article (the, ze, no)

BE (be)

BED (were)

BEDZ (was)

...

BEZ (is)

...

CD cardinal number (two, dozen, hundred...)

...

CD-CD hyphenated cardinal number (1985-1995...)

...

DO (do)

DOD (did)

...

HV (have)

...

IN preposition (after, by, of, for, since ...)

JJ general adjective (turquoise, happy...)

...

NN singular common noun (boy, pencil...)

NN\$ genitive singular common noun (boy's, parliament's ...)

...

RB adverb (else, about, hopefully ...)

...

TO infinitival TO

VB base form of lexical verb

VBD past tense of lexical verb

...

POS-taggelés a Brown-korpuszban

Television/NN has/HVZ yet/RB to/TO work/VB
out/RP a/AT living/RBG arrangement/NN with/IN
jazz/NN ,/, which/VDT comes/VBZ to/IN the/AT
medium/NN more/QL as/CS an/AT uneasy/JJ
guest/NN than/CS as/CS a/AT relaxed/VBN
member/NN of/IN the/AT family/NN ./.

<http://ilk.uvt.nl/~zavrel/tagtest.html>

Többértelműség a POS-taggelésben

The	AT		
man	NN	VB	
still	NN	VB	RB
saw	NN	VBD	
her	PPO	PP\$	

Egyértelműsítés rákövetkezés-gyakorisággal

The man still saw her.

The AT

man NN VB

still NN VB RB

saw NN VBD

her PPO PP\$

$$p(AT)=1.000$$

$$p(AT-NN)=0.186$$

$$p(AT-VB)=0.001$$

$$p(NN-NN)=0.040$$

$$p(NN-VB)=0.009$$

$$p(NN-RB)=0.040$$

$$p(AT-NN-NN)=p(AT-NN)*p(NN-NN)=0.07440$$

$$p(AT-NN-VB) = p(AT-NN)*p(NN-VB) = 0.01674$$

$$p(AT-NN-RB) = p(AT-NN)*p(NN-RB) = 0.07440$$

...

	NN	PPO	PP\$	RB	VB	VBD	.
AT	186	0	0	8	1	8	9
NN	40	1	3	40	9	66	186
PPO	7	3	16	164	109	16	313
PP\$	176	0	0	5	1	1	2
RB	5	3	16	164	109	16	313
VB	22	694	146	98	9	1	59
VBD	11	584	143	160	2	1	91

A Susanne-formátum

A01:0010b	AT	The	the	[O[S[Nns:s.
A01:0010c	NP1s	Fulton	Fulton	[Nns.
A01:0010d	NNL1cb	County	county	.Nns]
A01:0010e	JJ	Grand	grand	.
A01:0010f	NN1c	Jury	jury	.Nns:s]
A01:0010g	VVDv	said	say	[Vd.Vd]
A01:0010h	NPD1	Friday	Friday	[Nns:t.Nns:t]
A01:0010i	AT1	an	an	[Fn:o[Ns:s.
A01:0010j	NN1n	investigation	investigation	.
A01:0020a	IO	of	of	[Po.
A01:0020b	NP1t	Atlanta	Atlanta	[Ns[G[Nns.Nns]
A01:0020c	GG	+<apos>s	-	.G]
A01:0020d	JJ	recent	recent	.
A01:0020e	JJ	primary	primary	.
A01:0020f	NN1n	election	election	.Ns]Po]Ns:s]
A01:0020g	VVDv	produced	produce	[Vd.Vd]
A01:0020h	YIL	<ldquo>	-	.
A01:0020i	ATn	+no	no	[Ns:o.
A01:0020j	NN1u	evidence	evidence	.
A01:0020k	YIR	+<rdquo>	-	.
A01:0020m	CST	that	that	[Fn.
A01:0030a	DDy	any	any	[Np:s.
A01:0030b	NN2	irregularities	irregularity	.Np:s]

A Lancaster/IBM-formátum

B01000 1 v
[N The_AT world_NN1 N][V owes_VVZ [N a_AT1
considerable_JJ debt_NN1 [P of_IO [N gratitude_NN1
N]P]N][P[P&g to_II [N Mr._NNSB1 Byles_NP1 [N
the_AT Butcher_NN1 N]N]P&g] ,_. or_CC [P+ to_II [N
the_AT [less_RGR generic_JJ] alias_NN1 [Fr[P
under_II [N which_DDQ N]P][N this_DD1 indispensable_JJ
tradesman_NN1 N][V supplied_VVD [N the_AT needs_NN2
[P of_IO [N the_AT Stevenson_NP1 family_NN1 N]P]N]
[P in_II [N Bournemouth_NP1 N]P][P in_II [N 1885_MC
N]P]V]Fr]N]P+]P] ,_. [Fa since_ICS [N Stevenson_NP1 N]
[V explained_VVD [Fn that_CST [[N it_PPH1 N][V was_VBDZ
[N the_AT bills_NN2 [P of_IO [N Mr._NNSB1 Byles_NP1 N]
P]N]V]]][N which_DDQ N][V drove_VVD [N him_PPHO1 N]
[P to_II [N the_AT writing_NN1 [P of_IO [N The_AT
Strange_JJ Case_NN1 [P of_IO [N[N&g Dr._NNSB1 Jekyll_NP1
N&g] and_CC [N+ Mr._NNSB1 Hyde_NP1 ,_.
N+]N]P]N]P]N]P]V]]Fn]V]Fa]V]

A BNC-formátum

Eredeti:

Jordan lifts 22 years of martial law.

By Wafa Amr in Amman

JORDAN'S Prime Minister, Mr Mudar Badran, yesterday announced the freezing of martial law for the first time since 1967 as a prelude for formally revoking most of its provisions and abolishing it.

BNC-formátum:

```
<head type=BYLINE><s n=0002 p=Y><w II>By <w NP1>Wafa <w  
NP1>Amr <w II>in<w NP1>Amman </s></head>  
<p><s n=0003 p=Y><w NP1>JORDAN<w GE>'S <w JJ>Prime <w  
NN1>Minister<c YCOM>, <w NNB>Mr <w NP1>Mudar <w  
NP1>Badran<c YCOM>, <w RT>yesterday <w VVD>announced <w  
AT>the <w NN1>freezing <w IO>of <w JJ>martial <w  
NN1>law <w IF>for <w AT>the <w MD>first <w NNT1>time <w  
II>since <w MC>1967 <w II>as <w AT1>a <w N1>prelude <w  
IF>for <w RR>formally <w VVG>revoking <w DAT>most <w  
IO>of <w APPGE>its <w NN2>provisions <w CC>and <w  
VVG>abolishing <w PPH1>it<c YSTP>. </s></p>
```



Együtt-előfordulások a BNC-ban

turn British National Corpus freq = 44051

object	10205	2.3	subject	10255	4.6	modifier	7350	and/or	1947	0.3	part_trans	3430	15.4	
back	597	43.78	tide	71	30.29	0.6		twist	60	39.7	down	813	51.97	
attention	505	43.14	wheel	53	22.65	left	396	77.77	walk	108	36.3	over	500	51.14
head	752	39.77	Wycliffe	17	20.81	right	462	60.02	look	181	31.73	off	418	45.35
corner	232	36.86	Robyn	19	19.84	away	990	59.97	loss	29	31.07	round	169	41.3
page	178	31.08	Sheldukher	9	19.64	then	895	50.36	smile	36	26.82	out	469	36.74
handle	66	30.66	Rex	16	18.81	upside down	62	48.74	flee	23	26.35	up	489	35.98
somersault	19	29.8	Cardiff	21	18.11	now	517	41.73	stop	57	24.95	around	108	33.79
tap	56	29.33	Athelstan	14	17.42	inside out	26	41.02	run	74	24.3	on	210	31.56
knob	33	28.64	poacher	9	17.36	slowly	116	36.79	stare	25	22.46	back	191	30.26
key	97	27.02	Devore	7	16.64	half	52	35.93	shrug	15	21.69	in	44	15.39
eye	292	26.18	Luke	21	16.54	quickly	122	34.55	stride	11	21.1	about	10	8.83

A Penn Treebank

- ❑ Az első szintaktikailag elemzett korpusz
- ❑ Tartalma: kb. 3 millió szó
- ❑ Alapjai:
 - ❑ Brown-korpusz (Treebank I)
 - ❑ Wall Street Journal Corpus (Treebank II)
 - ❑ ATIS corpus
 - ❑ www.cis.upenn.edu/~treebank
 - ❑ tgrep-interfésszel:
www ldc.upenn.edu/ldc/online/treebank/

A Penn Treebank formátuma

[S[N Nemo_NP1 ,_, [N the_AT killer_NN1
whale_NN1 N] ,_, [Fr[N who_PNQS N][V
'd_VHD grown_VVN [J too_RG big_JJ [P
for_IF [N his_APP\$ pool_NN1 [P on_II [N
Clacton_NP1 Pier_NNL1
N]P]N]P]J]V]Fr]N] ,_, [V has_VHZ
arrived_VVN safely_RR [P at_II [N
his_APP\$ new_JJ home_NN1 [P in_II [N
Windsor_NP1 [I safari_NN1 park_NNL1
I]N]P]N]P]V] ._. S]

Néhány treebank-kereső

☐ Közvetlen összetevős szerkezetekre

VP << /[^]believe/ < (S < (/[^]NP/ !<< /[^{*}]/ !< (-NONE- < T)) < (VP|AUX << to))

(pl. tgrep, Penn Treebank)

(QP iDominates CONJ)

(pl. CorpusSearch, Penn, Helsinki)

☐ Függőségi szerkezetekre

[pos = "NN" & lemma = /Be.+ung/]

(pl. TigerSearch, Potsdam)

cat="NP" & #1:arity=4

(pl. ANNIS, Humboldt)

Grafikus keresők

(pl. PNL-TQ, Netgraph, Prága)



Magyar szövegtörzsek

Szeged Korpusz

- ☐ Az első magyar annotált korpusz
(SZTE + MTA NYTI + MorphoLogic közös projektje)
- ☐ szépirodalom (részletek Rejtő Jenő Pizkos Fred, a kapitány, Szerb Antal Utas és holdvilág, ill. George Orwell 1984 c. műveiből)
- ☐ 14-16 éves korú tanulók fogalmazásai
- ☐ újságcikkek (részletek a Népszabadság, a Népszava, a Magyar Hírlap, ill. a HVG egyes számaiból)
- ☐ számítógépes szövegek (részletek Kis Balázs Windows 2000 c. könyvéből, ill. a ComputerWorld/Számítástechnika magazin egyes számaiból)
- ☐ jogi szövegek (részletek a gazdasági társaságokról, ill. a szerzői jogokról szóló törvényekből)
- ☐ Morfológiailag egyértelműsített korpusz
 - ☐ Brill: 96,52%
 - ☐ TnT (HMM): 96,18%
 - ☐ RGLearn (szabály-alapú, saját): 94,54%
 - ☐ Kombinációjuk: 96,95%

Szeged Treebank

- ☐ Az első magyar treebank (kézi annotáció!)
 - ☐ határozói szerkezetek (ADVP)
 - ☐ jelzős szerkezetek (ADJP)
 - ☐ névutós szerkezetek (PP)
 - ☐ az igeneves szerkezetek (PA, INF)
 - ☐ tagadószók, igekötők, igék és kötőszavak
-
- ☐ A szabályok egy részét a konzorcium nyelvész szakértői készítették el és a kézzel definiált szakértői szabályokat az annotált treebankból számítógépes tanulási módszerekkel kinyert szabályokkal egészítették ki

A (nem címkézett) Magyar Webkorpusz

- ☐ Szószablya projekt (BME MOKK, 2003-2004)
- ☐ A .hu domén weboldalai
- ☐ 1,48 milliárd token (szűrés nélkül)
- ☐ Szűrések (az ismeretlen szavak százaléka):
 - 40%: gyakorlatilag nincs már benne nem magyar szöveg
 - 8%: gyakorlatilag nincs már benne ékezet nélküli szöveg
 - 4%: egy átlagszöveg elütéseinél kevesebb van benne
- ☐ A 4%-os letölthető: az 1220 millió magyar weboldal 589 millió szava tömörítve is 4 GB

	millió oldal	millió token	millió type
teljes	3,5	1486	19,1
40%	3,125	1310	15,4
8%	1,918	928	10,9
4%	1,221	589	7,2

A Magyar Nemzeti Szövegtár

- ❑ 1998 óta: <http://corpus.nytud.hu/mnsz/>
- ❑ A fele sajtó, benne van a DIA, és részben a MEK, továbbá index.hu-s fórumok szövegei és joganyagok
- ❑ 187,6 millió szó:

	magyaró.-i	szlovákiai	kárpátaljai	erdélyi	vajdasági	összesen
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

- ❑ Magyar Nemzeti Szövegtár 2: 500 millió szó



Az MNSZ szerinti szógyakoriság

Főnevek

...
30. év
...
54. ember
...
58. idő
...
62. rész
...
65. szó
...

Melléknevek

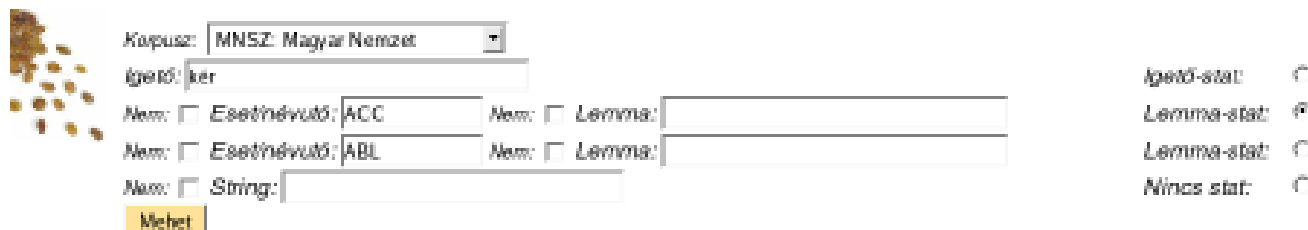
...
36. nagy
...
44. jó
...
46. magyar
...
62. új
...

Igék

szótő	szófaj	db
1. a	<i>Det</i>	11128421
2. az	<i>Det</i>	3716414
3. és	<i>Con</i>	2544751
4. hogy	<i>Con</i>	2166004
5. A	<i>Det</i>	2103970
6. az	<i>Pro</i>	1803814
7. nem	<i>Adv</i>	1693748
8. is	<i>Con</i>	1677108
9. van	<i>V</i>	1418113
10. ez	<i>Pro</i>	1204269
11. egy	<i>Num</i>	899832
12. Az	<i>Det</i>	730287
13. meg	<i>Pre</i>	592986
14. kell	<i>V</i>	499659
15. csak	<i>Adv</i>	477956
16. lesz	<i>V</i>	469189
17. de	<i>Con</i>	462508
18. már	<i>Adv</i>	452814
19. Ez	<i>Pro</i>	447310
20. amely	<i>Pro</i>	417945

Prószéky Gábor

A Mazsola vonzatkereső



Kopasz:

Irány:

Nem: ☐ Esetnévutó: Nem: ☐ Lemma:

Nem: ☐ Esetnévutó: Nem: ☐ Lemma:

Nem: ☐ String:

Igató-stat: ☐

Lemma-stat: ☐

Lemma-stat: ☐

Nincs stat: ☐

1010 találat. Stat: [bocsánat](#) [51] [segítség](#) [53] [elnézés](#) [32] [az](#) [136] [engedély](#) [32] [tájékoztatás](#) [21] [támogatás](#) [25] [péng](#) [20] [felmentés](#) [12] [válasz](#) [14] [tanács](#) [13] [íróint](#) [16] [magyarítás](#) [9] [igazolás](#) [8] [átírástól](#) [8] [kutatás](#) [7] [normaközlés](#) [6] [információ](#) [6] [teljesítmény](#) [7] [készenléti](#) [6] [megállapítás](#) [7] [készen](#) [6] [gazdaság](#) [6] [belsőkérdés](#) [6] az [12] [válasz](#) [7] [válasz](#) [6] az [7] [6]

adat

Az információk ellenőrzésére percase a kórháztól kér adatokat, Feladatai teljesítéséhez adatokat kérhet a bíróságtól, az ügyészségtől, a nemzetbiztonsági szolgálatoktól, a társadalombiztosítási igazgatási szervezettől. ha maximum öt főről kér adatot a hivataltól. hogy a rendőrség az adóhatóságtól, telefonaszolgálatoktól, bankoktól ügyései jóváhagyás nélkül kérjen adatokat, hogy kérje tőlük a művelődési tárca a bérfejlesztéshez szükséges adatokat.

adatgyűjtés

hanes Orbán Viktor és Deutsch Tamás ellen is adatgyűjtést kértek az ügyben érintett magányosozótól.

adóigazolás

amely szerint a gépkocsi átíratásakor adóigazolást kértek a polgároktól.

adókezelés

Korábban Budapest és a vidéki nagyvárosok külföldi adókezelésüket is kértek az eldől kabinettől,

adózás

Ha a bérbeadó magánszemély, akkor adószámlát kell kérnie az APEH-től,

ha a magánszemély adószámlát kér az APEH-től

aggregátor

A hadseregétől kértek aggregátort,

ajánlat

A helyreállítással megbízott Szabolcs-Szatmár-Bereg Megyei Közös Képzési Kht. tíz vállalkozótól másfél milliárd forintos felújítási munkára kért a közelmúltban kivitelezési ajánlat. Az eredménytelen pályázatok után a bizottság új ajánlatot kért a pályázóktól. és csatolva a pályázóktól - meghatározott műszaki paraméterekkel rendelkező - új kocsikra kért ajánlatot. hogy az MTK vezetői már ajánlatokat kértek néhány légitársaságtól a norvégiai repülőterrel kapcsolatban,

aláírás

A MÁV nyilatkozat aláírását kérte a szakaszvezetőktől,

hogy az aláírásokat lehetőleg olyan személyektől kell kérni.

alkotmányértelmezés

Az MDF alkotmányértelmezést kér a legfőbb ügyésztől

Az Országgyűlés alkotmányügyi bizottsága abstract alkotmányértelmezést kér az Alkotmánybíróságtól az ellenzék által kezdeményezett népszavazás ügyében.

így alkotmányértelmezést sem kérhet az Alkotmánybíróságtól.

Stílus-összehasonlítás a Mazsolával

PÉLDA

'ad -t' keret

a Magyar Nemzetben és az Index fórumban

8674 találat. hang [477] lehetőség [412] válasz [331] otthon [381] tájékoztatás [170] hír [309] ok [180] áldás [85]
esély [114] lendület [74] hely [150] pénz [129] magyarázat [87] szám [128] alkalom [101] tanács [85] koncert [71] munka [116]
utasítás [61] felvilágosítás [52] hangverseny [49] támogatás [93] kép [95] igaz [82] lökés [41] százalék [82] segítség [70]
engedély [56] megbízás [50] forint [85] információ [59] felhatalmazás [39] voks [38] ízelítő [29] alap [68] mód [59] jel [48] nyomaték [27] tér [46]

eltérések

7649 találat. igaz [30] válasz [275] tanács [172] hang [145] tipp [6] ami [212] lehetőség [155] pénz [151] magyarázat [88]
hála [67] ok [116] név [109] esély [93] ötlet [68] kegyelem [49] mi [118] hitel [61] vér [44] link [34] bizonyosság [31] hely [74] élet [66] cím [54]
felmentés [32] több [75] valami [59] fegyver [41] áldás [26] az [138] maga [62] alap [53] semmi [45] alkalom [40] fej [39] információ [37] jel [35] interjú [32]
parancs [28] megbízás [28] puszi [18] erő [40] hír [35] engedély [33] hit [30] kedvezmény [27] felvilágosítás [22] pofon [20] ez [74] sok [40] tér [34] kép [30]

Automatikus szemantikai osztályzás

nő

alany: szám arány ár

-bAn: mérték év

-rA: forint százalék dupla

-vAl: százalék forint

növekedik

alany: szám arány érdeklődés

-bAn: mérték év

-rA: év

-vAl: százalék forint

emelkedik

alany: szám ár árfolyam

-bAn: mérték év

-rA: forint jogerő százalék

-vAl: százalék forint

drágul

alany: részvény TVK kenyér

-bAn: mérték forgalom

-rA: forint

-vAl: százalék forint

Akár szótárgyártásra is használható

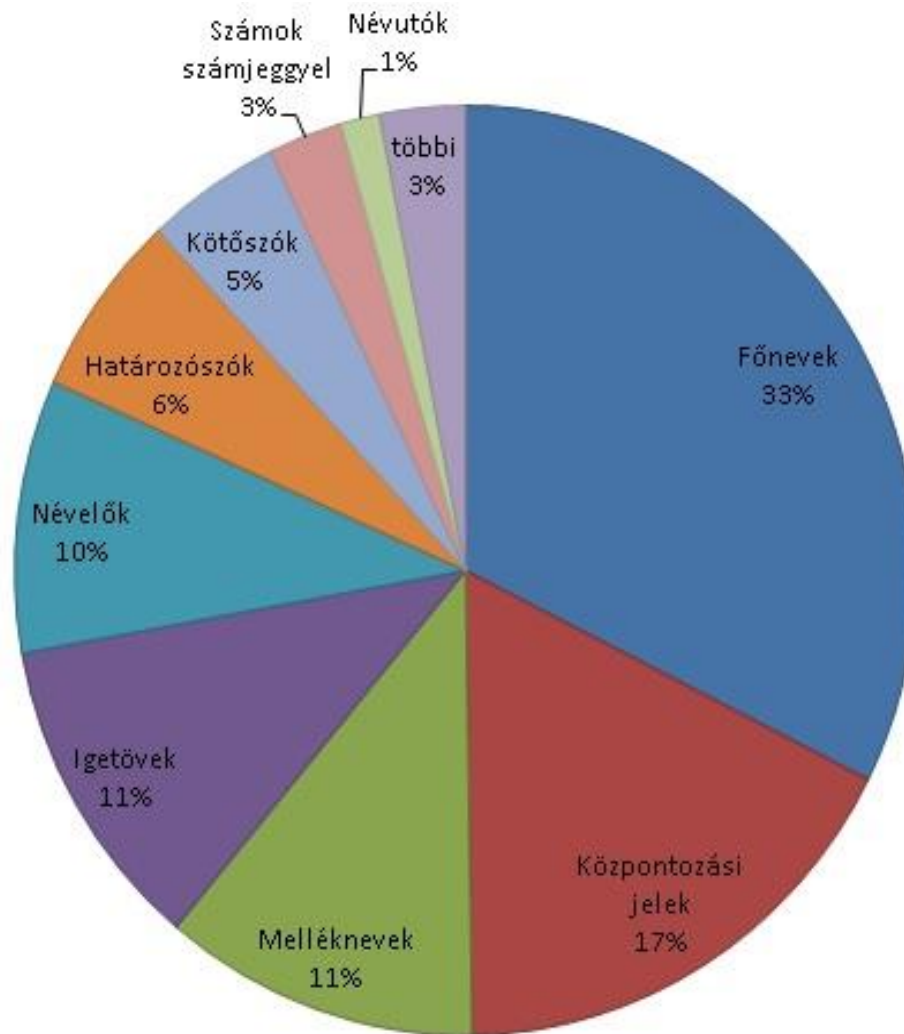
☐ -t hány	fitty
☐ -ra hány	szem
☐ alá kerül	víz, kalapács, fennhatóság
☐ alá rejt	véka
☐ alá hoz	tető
☐ alá helyez	vád
☐ alá vesz	górcső, kalap, tűz



A Pázmány Korpusz

alkorpusz	tokenszám	NP-szám	mondatszám
szztenderd szövegek	954 424 307	223 347 534	48 536 849
egyéb tartalmak	228 920 436	52 865 889	15 802 499
kommentek	59 013 104	13 867 066	3 505 818
összesen	1 242 357 847	290 080 489	67 845 166

A Pázmány Korpusz szófaji megoszlása





A Pázmány Korpusz szerkezete

eredeti szóalak	szótő	Humor- szófajcímke	módosított Humor-szófajcímke	főnévicsoport- határok
<s>				
Kezdetkor	kezdet	N+TEM	FN+TEM	1-N_1
teremtette	teremt	V+TMe3	IGE+TMe3	O
Isten	isten	N+NOM	FN+NOM	1-N_1
az	az	DET	DET	B-N_1
eget	ég	N+ACC	FN+ACC	E-N_1
és	és	CON	KOT	O
a	a	DET	DET	B-N_1
földet	föld	N+ACC	FN+ACC	E-N_1
.	.	PUNCT PERIOD	PUNCT	O
</s>				

A Pázmány Korpusz főnévi csoportjai

	A főnévi csoport felépítése	Korpuszbeli gyakoriság	Példa
1	[N+Nom]	14 312 895	ház
2	[Det] [N+Nom]	11 632 547	a ház
3	[N Pron+Nom]	6 416 317	ez
4	[Adj+Nom]	6 279 294	okos
5	[Det] [N+Acc]	4 616 117	a házat
6	[N Pron+Acc]	4 206 866	ezt
7	[N+Acc]	3 639 354	házat
8	[Det] [Adj] [N+Nom]	3 389 714	a nagy ház
9	[N+Nom] [N+Nom]	2 920 072	műanyag ház
10	[Det] [N+Pl+Nom]	2 690 635	a házak



A Pázmány Korpusz lekérdezhetősége

Query vár, FN.* 103,303 (83.2 per million)

[First](#) | [Previous](#) Page of 4,133 [Go](#) [Next](#) | [Last](#)

doc#5 normann lovaggal ostromolta meg Galeria várát , bevette , és foglyul ejtette Benedeket
doc#5 , VII. a Matild birtokában lévő Canossa várába ment , hogy várja a fejleményeket . 1077.
doc#5 25 : Henrik vezeklő ruhában jelent meg a vár kapuja előtt és 3 napon át kérte a perc.
doc#5 Pozsony ostrománál nagykorúsították , amikor a várat Borisztól visszavítta , IX. 11 : a Boriszt
doc#5 ostromolták . A háromszög alaprajzú hegyi várból jelentős romok állnak . Fügedi 1977:55
doc#5 vízhez , földalatti folyosót vágtak Sion várában , mely egy 13 m mély , függőleges aknához
doc#5 merítsenek . Dávid ezen keresztül vette be a várat : " Azon a napon Dávid azt mondta : ' Aki
doc#5 fölkapaszkodott a függőleges aknában (fönn a belső várban találta magát , s kinyitotta az ostromlók
doc#5 másztak meg , amikor bevették a jebuzeusok várat . Az akna felső végében egy ferdén felfelé
doc#5 karitatív intézete , neki adományozta Zólyom várat , s megbízta a Szepesség és a bányavárosok
doc#5 Hunyadi Giskraval , kinek kezén maradtak várai és jövedelmei . V. László megerősítette
doc#5 Mátyás csapatai mindenfelől támadták Giskra várai

Query vár, IGE.* 599,506 (482.6 per million)

[First](#) | [Previous](#) Page of 23,981 [Go](#) [Next](#) | [Last](#)

doc#7 ember keze nyújt , veszendő : | Halhatatlant vár s keres a valódi , | A nemes virtus : maga
doc#7 uralkodónéja . " A női közönségtől jótékony hatást vár a férfiúi nemre a hazai nyelv , az irodalom
doc#7 székvárosát , mert ebben az időben Napóleontól várta a monarchia és hazája feudális viszonyainak
doc#7 levegő árad ; olyasféle , amelyet már régen várt a kritika és a színház . A Rang és mód
doc#7 egyik grófi származású , rá nagy örökség vár . Sem a nagypapa , sem a két lány , se
doc#7 ellenezte az emigrációt is : tőle semmi jót nem várt , legfeljebb a Monarchia s vele együtt
doc#7 ugyanígy vélekedik : " Kisült , amit nem vártunk , | Mutatja a hatás : | Rossz drámaíró
doc#7 még hátra maradt , leküzdeni - ... mi vár tehát ránk ? Azt a jó isten tudja , ember
doc#7 fogantja lett . Nem volt más hátra , mint várni , és azt remélni , hogy az abszolutizmus
doc#7 kudarcra intő példa volt arra nézve , mi várna egy újabb magyar szabadságharcra . A kiegyezési
doc#7 az emberiség minden kérdésének megoldását várták . A harmincas , negyvenes évek nemzedékeinek
doc#7 mélységeket , melyeknél halál és kárhozat vár ránk ! " Martinuzzi alakjánál hangsúlyosabbnak



A nyelvtechnológia további területei

Néhány további kutatási terület

- ☐ Névelem-felismerés
- ☐ Anafora-felismerés
- ☐ Koreferencia-feloldás
- ☐ Témafelismerés
- ☐ Véleményelemzés
- ☐ Automatikus kivonatolás
- ☐ Diskurzuselemzés
- ☐ Természetesnyelv-generálás
- ☐ OCR-támogatás
- ☐ Kereséstámogatás
- ☐ Beszéd-alapú rendszerek támogatása
- ☐ Kérdés–válasz rendszerek
- ☐ ...

További információk a nyelvtechnológiáról

Association for Computational Linguistics

<http://www.aclweb.org>

ACL Video Archive

<http://www.fask.uni-mainz.de/lk/videoarchive>

ACL Anthology

<http://www.aclweb.org/anthology-new/>

DFKI

<http://registry.dfki.de>

PBML

<http://ufal.mff.cuni.cz/pbml.html>

META

<http://www.meta-net.eu/>

További információk, de már a vizsgáról

A végső osztályzat az alábbiakból alakul ki:

jelenlét az előadáson
(erre én emlékszem...)

+

aktív jelenlét a gyakorlaton
(ezt a gyakorlatvezetők tudják...)

+

házi feladatok megoldása
(ha volt ilyen, nyoma van...)

+

2 zárthelyi dolgozat eredménye
(a pontszámok adottak...)

+

vizsga az előadások anyagából
(amivel feljavítják a fentiek alapján kialakult osztályzatot...)

Vizsgatételek

1. Karakterek, kódolási szabványok, rendezések, n-gramok, karakter- és n-gram-alapú statisztikai megfigyelések
2. Véges állapotú automaták (és kiterjesztéseik) nyelvtechnológiai alkalmazásai
3. A számítógépes morfológia alapvető módszerei (kétszintes, unifikációs)
4. A számítógépes morfológia alkalmazásai (helyesírás, elválasztás, intelligens keresés, intelligens csere)
5. Formális nyelvtanok, szintaktikai reprezentációk (közvetlen összetevős, függőségi, X-vonás), mondatelemzési módszerek (TD, BU, kombinált)
6. Az unifikáció és alkalmazásai a nyelvtechnológiában (morfológiában, szintaxisban)
7. Jelentésreprezentációk, szemantikus hálók, fogalmi gráfok, ontológiák, WordNet
8. Számítógépes lexikográfia: szótárépítés, szótárátalakítás, szótárhasználat támogatása nyelvtechnológiával
9. Párhuzamos korpuszok, szövegszinkronizálás, fordítómemóriák
10. A gépi fordítás és módszerei