



# Korpuszok és adatbázisok

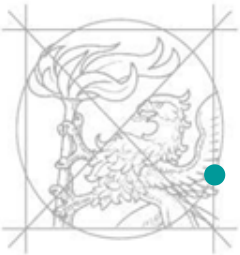


# Angol nyelvű korpuszok

- **British National Corpus (BNC)**
  - Brit angol
  - ~100M szövegszó
  - Írott és beszélt nyelv
  - Automatikus annotáció
- **Wall Street Journal (WSJ)**
  - Üzleti nyelv
  - Egyes részei kézzel annotálva (morfológia, szintaxis)
- **Reuters**
  - ~100 millió szövegszó
  - dokumentumok, bekezdések határai
- **Gigaword korpusz**
  - 2 milliárd szó
- **Penn TreeBank**
  - 5 millió szövegszó
  - szófaji kód
  - szintaktikai elemzés (konstituensfa)
- **Feladatspecifikus korpuszok: CoNLL-2003 (tulajdonnevek), SemEval (szemantika)...**
  - néhány százezer szövegszó

# Magyar Nemzeti Szövegtár (MNSZ)

- 187,6 millió szövegszó
- Sajtó, szépirodalom, tudományos, hivatalos, személyes szövegek
- Határon túli nyelvváltozatok is
- Automatikus szótövezés és szófaji elemzés
- Gigaword verzió (1 milliárd szövegszó)
- <http://corpus.nytud.hu/mnsz>



# Webkorpusz

- több mint 1,48 milliárd szó (szűretlenül, illetve 589 millió megszűrt szó)
- jelenleg a legnagyobb magyar nyelvű korpusz
- 18 millió weboldal (.hu)
- <http://mokk.bme.hu/resources/webcorpus>



# Párhuzamos korpuszok

- olyan két- vagy többnyelvű korpuszok, amelyben egy mű és annak egy vagy több nyelvre lefordított változatai szerepelnek
- Bekezdés-, mondat- vagy szószinten párhuzamosított szövegek
- Alkalmazási lehetőségek: fordítástudomány, kontrasztív nyelvészet, gépi fordítás...



# Néhány párhuzamos korpusz

- **Hansard**: angol-francia
- **1984**: közép- és kelet-európai nyelvek
- **Hunglish**: magyar-angol
- **SzegedParalell**: magyar-angol
- **HunOr**: magyar-orosz



# Szegedi korpuszok

- Szeged Treebank
- Szeged Dependencia Treebank
- Magyar WordNet
- Bizonytalanságra annotált korpuszok
- Tulajdonnévkorpuszok
- Lemmatizált tulajdonnevek
- Többszavas kifejezések korpuszai
- Jelentés-egyértelműsített korpusz
- Kutatói adatok HTML korpusza
- SzegedParalell
- HunOr
- Véleménydetekciós korpusz
- Kulcsszókinyerési korpuszok
- HunLearner

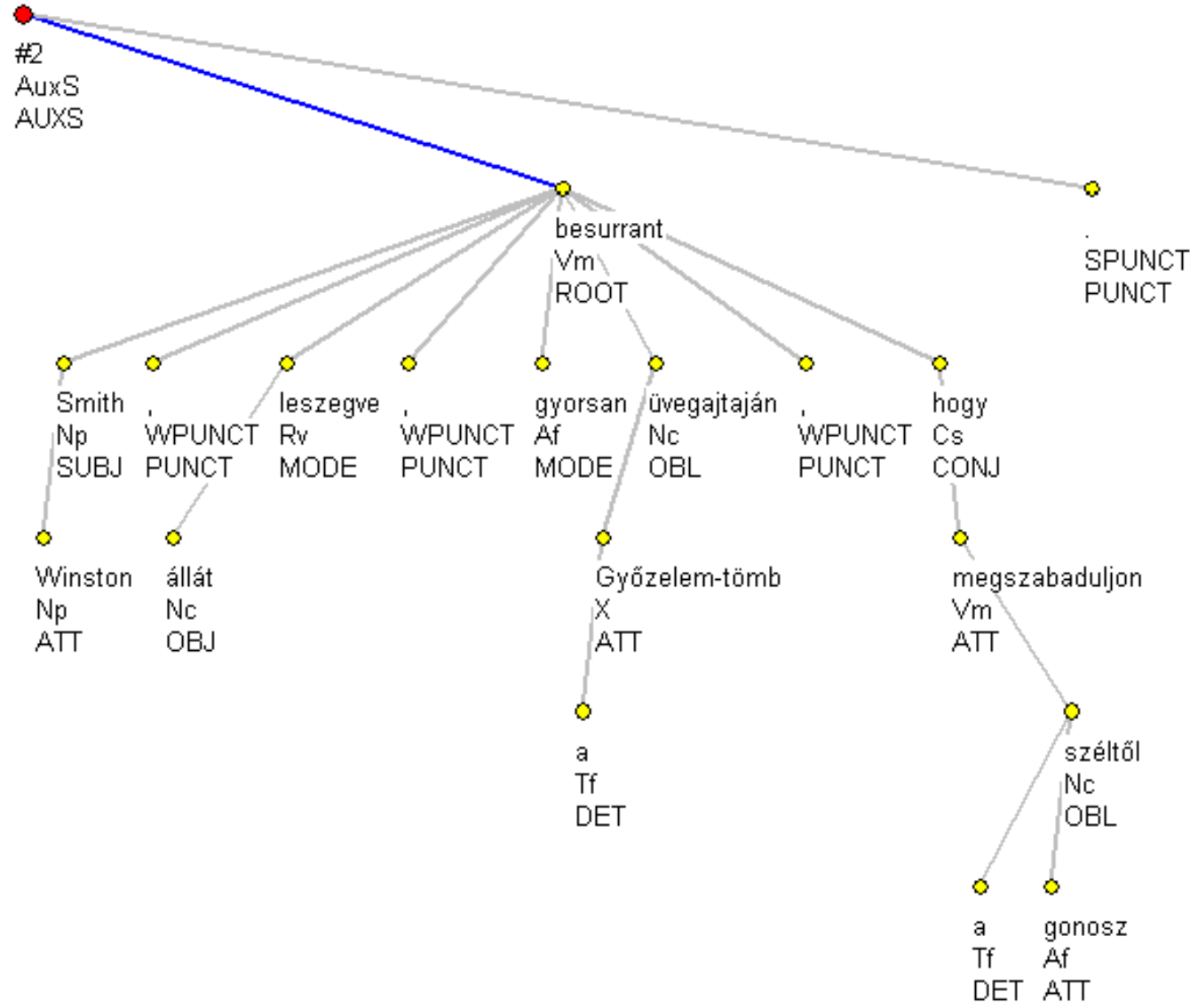
[http://www.inf.u-szeged.hu/rgai/nlp\\_download](http://www.inf.u-szeged.hu/rgai/nlp_download)

# Szeged (Dependencia) Treebank

- 82 000 mondat
- 1,5 millió szövegszó
- 230 000 írásjel
- 6 domén
  - iskolai fogalmazások
  - számítógépes szövegek
  - irodalom
  - jogi szövegek
  - újságcikkek
  - üzleti rövidhírek
- Kézzel ellenőrzött morfológiai és szintaktikai (konstituens és függőségi) elemzés, névelemek, félig kompozicionális szerkezetek (FX), koreferencia
- <http://www.inf.u-szeged.hu/rgai/SzegedTreebank>

|    |           |                                                            |       |     |     |                                           |         |     |     |
|----|-----------|------------------------------------------------------------|-------|-----|-----|-------------------------------------------|---------|-----|-----|
| 1  | –         | ROOT                                                       | ROOT  | ELL | ELL | –                                         | –       | 0   | 0   |
| 2  | Japánban  | Japán                                                      | Japán | N   | N   |                                           |         |     |     |
|    |           | SubPOS=p Num=s Cas=2 NumP=none PerP=none NumPd=none        |       |     |     |                                           |         |     |     |
|    |           | SubPOS=p Num=s Cas=2 NumP=none PerP=none NumPd=none        |       |     |     | 1                                         | 1       | OBL |     |
|    |           | OBL                                                        |       |     |     |                                           |         |     |     |
| 3  | ,         | PUNCT                                                      | PUNCT | ,   | ,   | –                                         | –       | 1   | 1   |
| 4  | ahol      | ahol                                                       | R     | R   |     | SubPOS=r Deg=none Num=none Per=none       |         |     |     |
|    |           | SubPOS=r Deg=none Num=none Per=none                        | 9     | 9   |     | TLOCY                                     | TLOCY   |     |     |
| 5  | 1960-ban  | 1960                                                       | M     | M   |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none |       |     |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none |       |     |     | 9                                         | 9       | OBL |     |
|    |           | OBL                                                        |       |     |     |                                           |         |     |     |
| 6  | közel     | közel                                                      | R     | R   |     | SubPOS=x Deg=none Num=none Per=none       |         |     |     |
|    |           | SubPOS=x Deg=none Num=none Per=none                        | 7     | 7   |     | MODE                                      | MODE    |     |     |
| 7  | félmillió | félmillió                                                  | M     | M   |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=n Form=l NumP=none PerP=none NumPd=none |       |     |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=n Form=l NumP=none PerP=none NumPd=none |       |     |     | 8                                         | 8       | ATT |     |
|    |           | ATT                                                        |       |     |     |                                           |         |     |     |
| 8  | válást    | válás                                                      | N     | N   |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=a NumP=none PerP=none NumPd=none        |       |     |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=a NumP=none PerP=none NumPd=none        |       |     |     | 9                                         | 9       | OBJ |     |
|    |           | OBJ                                                        |       |     |     |                                           |         |     |     |
| 9  | mondtak   | mond                                                       | V     | V   |     | SubPOS=m Mood=i Tense=s Per=3 Num=p Def=n |         |     |     |
|    |           | SubPOS=m Mood=i Tense=s Per=3 Num=p Def=n                  |       |     |     | 1                                         | 1       | ATT | ATT |
| 10 | ki        | ki                                                         | R     | R   |     | SubPOS=p Deg=none Num=none Per=none       |         |     |     |
|    |           | SubPOS=p Deg=none Num=none Per=none                        | 9     | 9   |     | PREVERB                                   | PREVERB |     |     |
| 11 | ,         | PUNCT                                                      | PUNCT | ,   | ,   | –                                         | –       | 9   | 9   |
| 12 | 1990-ben  | 1990                                                       | M     | M   |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none |       |     |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=2 Form=d NumP=none PerP=none NumPd=none |       |     |     | 1                                         | 1       | OBL |     |
|    |           | OBL                                                        |       |     |     |                                           |         |     |     |
| 13 | már       | már                                                        | R     | R   |     | SubPOS=x Deg=none Num=none Per=none       |         |     |     |
|    |           | SubPOS=x Deg=none Num=none Per=none                        | 15    | 15  |     | MODE                                      | MODE    |     |     |
| 14 | 2,6       | 2,6                                                        | M     | M   |     |                                           |         |     |     |
|    |           | SubPOS=f Num=s Cas=n Form=d NumP=none PerP=none NumPd=none |       |     |     |                                           |         |     |     |
|    |           | SubPOS=f Num=s Cas=n Form=d NumP=none PerP=none NumPd=none |       |     |     | 15                                        | 15      | NUM |     |
|    |           | NUM                                                        |       |     |     |                                           |         |     |     |
| 15 | milliót   | millió                                                     | M     | M   |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=a Form=l NumP=none PerP=none NumPd=none |       |     |     |                                           |         |     |     |
|    |           | SubPOS=c Num=s Cas=a Form=l NumP=none PerP=none NumPd=none |       |     |     | 1                                         | 1       | OBJ |     |
|    |           | OBJ                                                        |       |     |     |                                           |         |     |     |
| 16 | .         | PUNCT                                                      | PUNCT | .   | .   | –                                         | –       | 0   | 0   |





# WordNet

- **Lexikális adatbázis**
- **Fogalmak hálóba rendezve  
különféle relációk alapján**
- **Angol: Princeton WordNet (PWN)**
- **Más nyelvekre is: EuroWordNet,  
BalkaNet stb.**
- **Magyar: Hungarian WordNet  
(HuWN)**



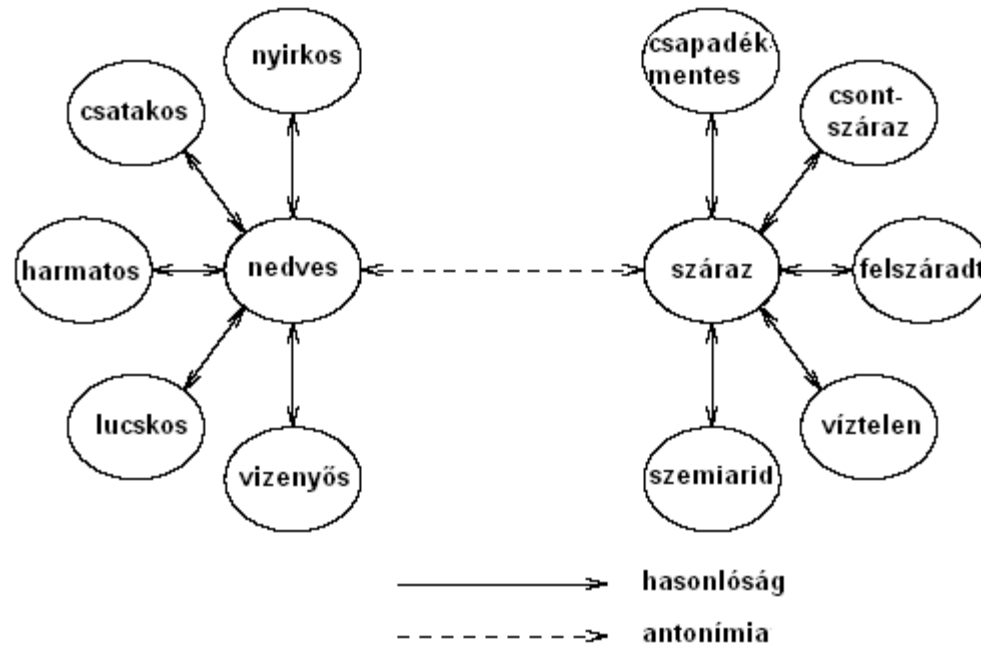
# A HuWN bemutatása

- 40 000 synset (általános ontológia)  
+ 2000 üzleti nyelvi, ill. 650 jogi  
nyelvi synset (szakontológia)
- Főnevek
- Igék
- Melléknevek
- Határozószók
- Alapelv: ahol csak lehet, a PWN-  
nek megfeleltetni a synseteket

<http://www.inf.u-szeged.hu/rgai/HuWN>



# Melléknevek a HuWN-ben



# WSD korpusz

- Jelentés-egyértelműsítés
- A WordNet építése mellett elkészült Szegeden az első (Lexical Sample) tanítókörpusz magyarra (finom jelentésmegkülönböztetés)
- 39 szóalak
- szóalakonként 300-500 címkézett példa
- 6 melléknév:  
*anyagi, élő, erős, képes, pontos, szociális*
- 21 főnév:  
*civil, család, élet, ház, helyzet, intézmény, iskola, kép, képviselő, kormány, nap, oldal, ország, perc, pont, program, század, személy, szervezet, tanár, világ, víz*
- 12 ige:  
*függ, hat, jár, kap, kerül, marad, rendelkezik, szerepel, tart, tartozik, tud, válik*

[http://www.inf.u-szeged.hu/rgai/corpus\\_hunwsd](http://www.inf.u-szeged.hu/rgai/corpus_hunwsd)



**SenseTagger - család\_n**

File View Help

**Context**

összefüggéseket hangsúlyoz , valójában annak a romboló hatásnak a következménye , amelyet a marxizmus gyakorolt a szellemi életre .  
 Pedig a sem marxistának , sem balosnak nem tekinthető Charles Darwin Az ember származása című művében maga is azt írja , hogy egy család , nemzetség vagy faj túlélési esélyei annál nagyobbak lehetnek , minél nagyobb mértékben képesek az egyedei a részvételre , az együttműködő és önzetlen magatartásra .  
 Az önzés mint az evolúció hajtóereje éppen ezért nem Darwintól származó , hanem neodarwinista gondolat - véli Az arccal a másik felé című könyv harvardi szerzőpárosa .  
 De hát akkor valójában az önzés avagy az önzetlenség evolúciójáról beszélhetünk inkább ?  
 Annyi bizonyosnak látszik - fejtik ki a szerzők - , hogy egy nemzedékkel ezelőtt a biológushallgatók

Current sense id: család\_n\_2\_rendszertani 29 / 400

<<< << >> >>>

**Senses**

| Sense id               | Definition          | N... |
|------------------------|---------------------|------|
| (none)                 | No classification   | 0    |
| csalad_n_6_egye1b      | egyéb               | 18   |
| csalad_n_1_rokonsa1g   | rokonság            | 370  |
| csalad_n_2_rendszer... | rendszertani egység | 2    |
| csalad_n_3_gyerek      | gyerek népiesen     | 0    |
| csalad_n_4_uralkodo... | uralkodóház         | 10   |
| csalad_n_3_me1hcsal... | méhcsalád           | 0    |

Add

Modify

Remove

**Definition:**

rendszertani egység

Exit

# NE-korpuszok

- CoNLL-verseny normáit követi
- ORG / LOC / PER / MISC osztályok
- ~220 000 szövegszó (SZK üzleti hírek)
- ~470 000 szövegszó (HVG-cikkek)
  - Szó szerinti (tag-for-tag)
  - Metonimikus jelölés (tag-for-meaning)

[http://www.inf.u-szeged.hu/rgai/corpus\\_ne](http://www.inf.u-szeged.hu/rgai/corpus_ne)



# SzegedParalell

- Magyar-angol párhuzamos korpusz
- Kézzel párhuzamosított bekezdés és mondat szinten:
  - nyelvkönyvek
  - EU-s szövegek
  - Kétnyelvű újságok
  - irodalom
- 99.000 mondat szintű egység
- Egy része FX-ekre annotálva

[http://www.inf.u-szeged.hu/rgai/corpus\\_paralell](http://www.inf.u-szeged.hu/rgai/corpus_paralell)

# Bizonytalanságra annotált korpuszok

- BioScope (20K mondat)
  - Orvosi szövegek
  - Biológiai absztraktok
  - Biológiai cikkek
- CoNLL-2010 Shared Task korpuszok (Biológiai cikkek (18K mondat) + Wikipedia-szócikkek (20K mondat) )
- Szeged Uncertainty Corpus
  - Újraannotált CoNLL-2010 + FactBank
  - Egységes annotációs elvek
- WikiWeasel 2.0: diskurzusszintű bizonytalanság
- hUnCertainty: magyar korpusz (17K mondat)

<http://www.inf.u-szeged.hu/rgai/uncertainty>

A O O

lap O O

szerint B-doxastic B-doxastic

P. O O

. O O

Márió O O

kitart B-doxastic O

amellet O O

, O O

hogy O O

egyáltalán O O

nem O O

emlékszik O O

arra O O

, O O

hogy O O

őt O O

bárki O O

is O O

üldözte O O

volna O O

. O O

Állítólag B-epistemic B-epistemic

azon O O

a O O

területen O O

, O O

ahol O O

a O O

vérengzés O O

történt O O

, O O

csak O O

a O O

gyilkos O O

kocsijának O O

a O O

keréknyomát O O

találták O O

meg O O

# MWE-korpuszok

- Többszavas kifejezések
- Wiki50 korpusz:
  - 50 angol Wikipedia-szócikk (4700 mondat)
  - MWE-k és NE-k kézzel jelölve
- Szeged Treebankben és SzegedParalell egy részében FX-ek
- JRC-Acquis jogi párhuzamos korpuszban FX-ek angol, német, spanyol és magyar nyelven (~100K token minden nyelven)

<http://www.inf.u-szeged.hu/rgai/mwe>

# Wiki50

On August 20, 2005, in a private ceremony, Thompson's ashes were fired from a cannon atop a tower of his own design (in the shape of a double-thumbed fist clutching a peyote button) to the tune of Norman Greenbaum's "Spirit in the Sky" and Bob Dylan's Mr. Tambourine Man. Red, white, blue, and green fireworks were launched along with his ashes. As the city of Aspen would not allow the cannon to remain for more than a month, the cannon has been dismantled and put into storage until a suitable permanent location can be found. According to his widow Anita, Thompson's funeral was financed by actor Johnny Depp, a close friend of Thompson. Depp told the Associated Press, "All I'm doing is trying to make sure his last wish comes true. I just want to send my pal out the way he wants to go out." Other famous attendees at the funeral included U.S. Senator John Kerry and former U.S. Senator George McGovern; 60 Minutes correspondent Ed Bradley and Charlie Rose; actors Jack Nicholson, Bill Murray, Benicio del Toro, Sean Penn, and Josh Hartnett; singers Lyle Lovett, John Oates and numerous other friends. An estimated 280 people attended the funeral.

The plans for this monument were initially drawn by Thompson and Ralph Steadman and were shown as part of an Omnibus program on the BBC entitled Fear and Loathing in Gonzo (1978). It is included as a special feature on the second disc of the 2003 Criterion Collection DVD release of Fear and Loathing in Las Vegas (labeled on the DVD as "Fear and Loathing on the Road to Hollywood"). The video footage of Steadman and Thompson drawing the plans and outdoor footage showing where he wanted the cannon constructed were planned prior to the unveiling of his cannon at the funeral.



# HunLearner

- Középhaladó és haladó szintű tanulók fogalmazásai
- Többségében horvát vagy észt anyanyelvűek
- Számítógépen, szótár és nyelvkönyv nélkül írt fogalmazások
- 1400 mondat
- Főnévi morfológiai hibák jelölve
- Alanyi/tárgyas ragozási hibák

<http://www.inf.u-szeged.hu/rqai/hunlearner>

|    |              |              |         |                 |             |   |                                                           |        |
|----|--------------|--------------|---------|-----------------|-------------|---|-----------------------------------------------------------|--------|
| 1  | A            | a            | Tf      | 2               | DET         | T | SubPOS=f                                                  |        |
| 2  | gyerek       | gyerek       | Nc-sn   | 9               | SUBJ        | N |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=c Num=s Cas=n NumP=none PerP=none NumPd=none       |        |
| 3  | nagyon       | nagyon       | Rx      | 4               | MODE        | R |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=x Deg=none                                         |        |
| 4  | okos         | okos         | Afp-sn  | 9               | ATT         | A |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=f Deg=p Num=s Cas=n NumP=none PerP=none NumPd=none |        |
| 5  | és           | és           | Ccsw    | 4               | CONJ        | C |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=c Form=s Coord=w                                   |        |
| 6  | kedves       | kedves       | Afp-sn  | 5               | COORD       | A |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=f Deg=p Num=s Cas=n NumP=none PerP=none NumPd=none |        |
| 7  | és           | és           | Ccsw    | 6               | CONJ        | C |                                                           |        |
|    |              |              |         |                 |             |   | SubPOS=c Form=s Coord=w                                   |        |
| 8  | jól          | jól          | Rxp     | 7               | COORD       | R | SubPOS=x Deg=p                                            |        |
| 9  | működik      | működik      | X       | 0               | ROOT        | X | —                                                         |        |
| 10 | a            | a            | Tf      | 11              | DET         | T | SubPOS=f                                                  |        |
| 11 | kapcsolatünk | kapcsolatünk | X       | 9               | OBL         | X |                                                           |        |
|    | kapcsolatunk |              | Stem: A | Assimilation: 1 | Matching: B |   | —                                                         | Suffix |
|    | number: 1    |              |         |                 |             |   |                                                           |        |
| 12 | .            | .            | .       | 0               | PUNCT       | . |                                                           |        |

# Véleménydetekciós korpusz

- Népszavazás a kettős állampolgárságról
- 1294 fórumhozzászólás
- *Igennel/nemmel szavazna – érvénytelenül szavaz – nem releváns* kategóriák szerint felcímkézve

[http://www.inf.u-szeged.hu/rgai/corpus\\_forum](http://www.inf.u-szeged.hu/rgai/corpus_forum)



# Személyiségjegyek és vélemények

- 500 utazási blog 5 úticélhoz kapcsolódva
- Angol nyelvű
- Pozitív és negatív vélemények adott dologra vonatkoztatva
- Személyiségjegyekre utaló szövegrészek is jelölve





|          | Barcelona     | Hungary     | India       | Los Angeles   | Middle East   | Total         |
|----------|---------------|-------------|-------------|---------------|---------------|---------------|
| Op       | 988           | 831         | 850         | 930           | 882           | 4,481         |
| PosOp    | 821 (83.10)   | 706 (84.96) | 632 (74.35) | 801 (86.13)   | 738 (83.67)   | 3698 (82.53)  |
| NegOp    | 167 (16.90)   | 125 (15.04) | 218 (25.65) | 129 (13.87)   | 144 (16.33)   | 783 (17.47)   |
| Desc     | 1,478         | 1,148       | 1,256       | 1,226         | 1,251         | 6,359         |
| PosDesc  | 1,241 (83.96) | 965 (84.06) | 921 (73.33) | 1,064 (86.79) | 1,047 (83.69) | 5,238 (82.37) |
| NegDesc  | 237 (16.04)   | 183 (15.94) | 335 (26.67) | 162 (13.21)   | 204 (16.31)   | 1,121 (17.63) |
| Target   | 947           | 806         | 829         | 892           | 841           | 4,315         |
| PersMark | 358           | 250         | 308         | 235           | 315           | 1,466         |
| Sentence | 4,152         | 3,644       | 3,769       | 4,170         | 3,926         | 19,661        |
| Token    | 87,624        | 79,386      | 76,533      | 83,161        | 83,266        | 409,970       |

Table 1: Statistical data on the annotated categories. Op: opinion, Desc: descriptor, Pos: positive, Neg: negative, PersMark: personality marker.

|                   | Barcelona | Hungary | India | Los Angeles | Middle East | Total  |
|-------------------|-----------|---------|-------|-------------|-------------|--------|
| Venturer          | 80        | 88      | 75    | 44          | 68          | 355    |
| Dependable        | 149       | 95      | 104   | 96          | 133         | 577    |
| Extraversion      | 55        | 17/1    | 15/7  | 20/2        | 26/3        | 133/13 |
| Agreeableness     | 3         | 3       | 11    | 3           | 3           | 23     |
| Openness          | 10        | 15      | 16    | 8           | 23          | 72     |
| Conscientiousness | 10/5      | 5/4     | 8/4   | 2/2         | 4/3         | 29/18  |
| Neuroticism       | 1         | 0       | 0     | 0           | 2/1         | 3/1    |

Table 2: Statistical data on personality markers.

# Koreferenciakorpusz

- Azonos referenciájú elemek összekötése
- Szeged Treebank szövegei
- Folyamatban...

Egy szép napon kirándulni ment a család . Első úti célunk a parádfürdői *Kocsi Múzeum* meglátogatása volt . Nagyon érdekesnek találtam a régi korok posta — és személyszállító kocsiait . *Innen* átmentünk a *lovardába* . Szép *méneket* láttunk . Parádsasvár , *Mátraháza* érintésével megérkeztünk *Kékestetőre* . *Itt az ország legmagasabb pontján meglátogattuk a TV-tornyot* . *Innen* pár órát eltöltve , indultunk tovább , megérkeztünk a Sástóhoz . A barátnőmmel kibéreltünk egy *csónakot* . A tavon evezve *vadrécéket* láttunk , *melyek* oly szelídek voltak , hogy a tenyerünkől lehetett *őket* etetni . A *csónakázás* befejeztével nekifogtunk *nyársalni* . Közben jött egy nyári zápor , és bőrig áztunk . Újra elkezdtek a *nyársaláshoz* az előkészületeket . A szüleim itt bebizonyították proOBJ , hogy nem indiánok . Kis proPOSS híján az ott üdülő család majdhogynem a tűzoltókat értesítette , mert olyan füstöt eresztett az általunk rakott tűz , hogy azt proSUBJ hitték , ég a *Mátra* . Végül mégis csak sikerült *nyársalni* . Az igaz , hogy úgy nézett ki a *szalonna* , mint a krampusz . Mégis elfogyasztottuk proOBJ . Ezután a szükséges tűzvédelmi feladatot elvégezve , vigyázva arra , hogy a környezetet ne szennyezzük , összeszedtük a szemetet . Majd kevés séta után úgy döntöttünk , hogy felmegyünk a *kilátóba* . *Ahonnan* meglepetésemre csodaszép *kilátás* nyílik a Mátra völgyére . A *kilátóból* lejövet egy kis fagyizás után , pihenőre tértünk . Majd úgy döntöttünk , hogy *horgászni* fogunk . Nekikezdünk a *horgászatok előkészítéséhez* . *Ez* meg is történt , egy fél kenyér begyúrása után . Mert *én* *magam* azon voltam , ha netán mégis sikerül egy *halat megfogni* , *én* jól megetetem őket . Sajnos *kapásom* az volt , de a türelmem elfogyott a sok szűnyog miatt . *Ezután* elmentünk egy *étterembe* vacsorázni . A sok mászkálás és a jó levegő megtette a proPOSS hatását . Egyszerűen nem akartam proOBJ elhinni , jó hogy ennyi *ételt* és ilyen jó étvággal el tudok fogyasztani . *Innen* elindultunk a szállásra pihenni . Ilyen jól ritkán alszom , annak ellenére , hogy nem a megszokott helyemen töltöttem az éjszakát . Másnap elindultunk haza .