# A HG-1 Treebank és keresőfelület fejlesztői munkái, használata és felhasználhatósága

Az elemzésektől a keresőfelületig



Csernyi Gábor

# 1 Célok, előzmények

Mit? (Mi a cél?)

1,5 millió szavas, elemzett mondatokból álló adatbázist  létrehozni

Miből?

a Magyar Webkorpuszból (Kornai et al., 2006)

Mivel?

XLE-vel, LFG-nyelvtannal (morfológia+lexikon+szintaxis)

# 2 A korpusz létrehozása

# 2 A korpusz létrehozása (2)

A Magyar Webkorpusz – mint forráskorpusz – mondatainak elemeztetése szuperszámítógép segítségével (HPC):

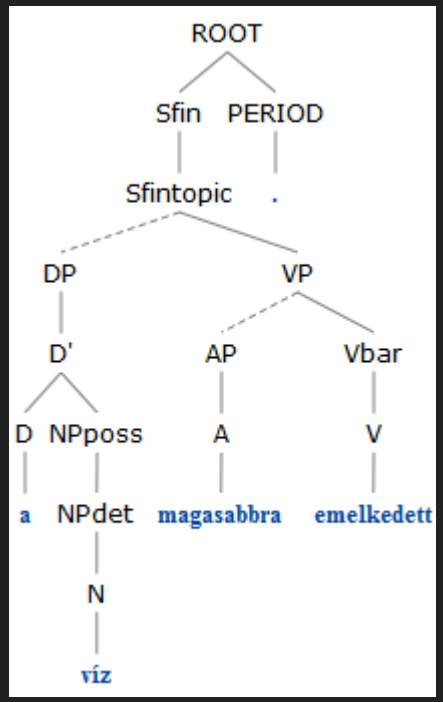
- kötegelt elemeztetés;
- párhuzamosított elemeztetés a forráskorpusz darabolásával;
- a 2-nél több elemzéssel rendelkező mondatok „eldobása”.

# 2 A korpusz létrehozása (3)

Az elemzések összegyűjtése, formátumra vonatkozó kérdések:

- XLE kimeneti formátuma
  - vizualizált reprezentáció helyett szöveges állomány (Prolog reprezentáció);

# 2 A korpusz létrehozása (3)




```
Fájl  Szerkesztés  Beállítások  Kikódolás  Súgó                    25 %
% -*- coding: utf-8 -*-

fstructure('A víz magasabbra emelkedett.',
        % Properties:
        [
        'xle_version'('XLE release of Nov 20, 2008 14:00.'),
        'grammar'('/home/XLE/hungram-111024/hungram1.lfg'),
        'grammar_date'('Oct 17, 2011 16:35'),
        'word_count'('4'),
        'statistics'('1 solutions, 0.01 CPU seconds, 55 subtrees unified'),
        'rootcategory'('ROOT'),
        'hostname'('linux-id85')
        ],
        % Choices:
        [

        ],
        % Equivalences:
        [

        ],
        % Constraints:
        [
        cf(1,eq(attr(var(0),'PRED'),semform('emelkedik',2,[var(3)],[]))),
        cf(1,eq(attr(var(0),'SUBJ'),var(3))),
        cf(1,eq(attr(var(0),'ADJUNCT'),var(1))),
        cf(1,eq(attr(var(0),'TOPIC'),var(7))),
        cf(1,eq(attr(var(0),'FOCUS'),var(2))),
        cf(1,eq(attr(var(0),'TNS-ASP'),var(6))),
        cf(1,eq(attr(var(3),'PRED'),semform('víz',0,[],[]))),
        cf(1,eq(attr(var(3),'NTYPE'),var(4))),
        cf(1,eq(attr(var(3),'CASE'),'nom')),
        cf(1,eq(attr(var(3),'DEF'),'+')),
        cf(1,eq(attr(var(3),'NUM'),'sg')),
        cf(1,eq(attr(var(3),'PERS'),'3')),
        cf(1,eq(attr(var(4),'NSEM'),var(5))),
        cf(1,eq(attr(var(4),'NSYN'),'common')),
        cf(1,eq(attr(var(5),'COMMON'),'+')),
        cf(1,in_set(var(2),var(1))),
        cf(1,eq(attr(var(2),'PRED'),semform('magas',1,[],[]))),
        cf(1,eq(attr(var(2),'CASE'),'sublative')),
        cf(1,eq(attr(var(2),'DEGREE'),'comparative')),
        cf(1,eq(attr(var(2),'NUM'),'sg')),
```

Nyers korpusz kiválasztása, majd elemeztetése (XLE) | Elemző (XLE) kimenetének feldolgozása | Adatbázis összeállítása, konverziók | Keresőfelület

# 2 A korpusz létrehozása (3)

Az elemzések összegyűjtése, formátumra vonatkozó kérdések:

- XLE kimeneti formátuma
  - vizualizált reprezentáció helyett szöveges állomány (Prolog reprezentáció);
  - minden egyes mondat elemzése(i) egy-egy fájlban.

# 2 A korpusz létrehozása (4)

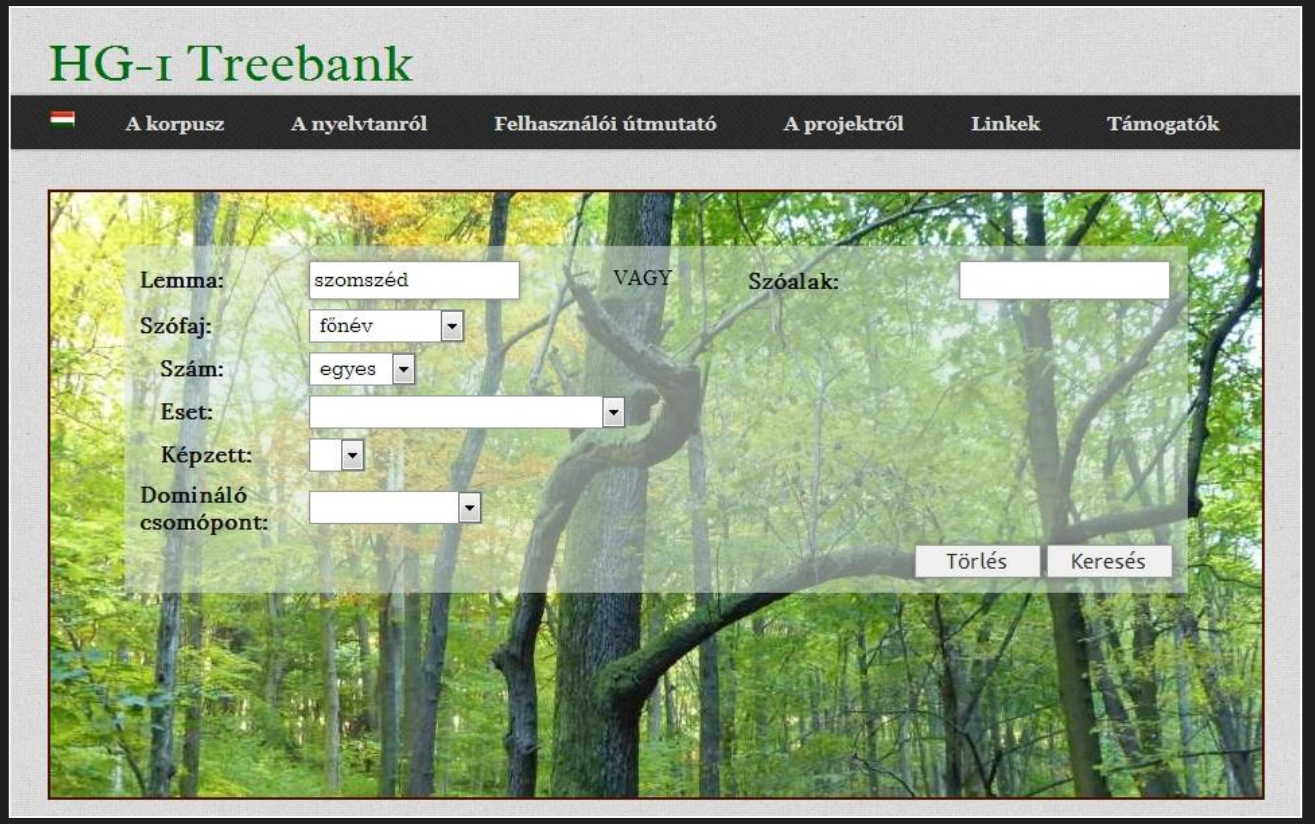
- Az egyes mondatelemzések (mondatonként egy-egy fájl) összefűzése;
- azonos elemzések kiszűrése;
- statisztikai információk gyűjtése az elemzésekről (pl. ágrajzok mérete);
- elemzések XML (TigerXML) adatbázisba rögzítése, konverziók (a későbbi vizualizációhoz).

# 2 A korpusz létrehozása (5)

Relációs adatbázis (XML-ből konvertálva)

„Dinamikus (PHP-alapú) rendszer

Keresés beállítása űrlap segítségével



# 3 Az online keresőfelület

**On-line lekérdezési felület főbb funkciói:**

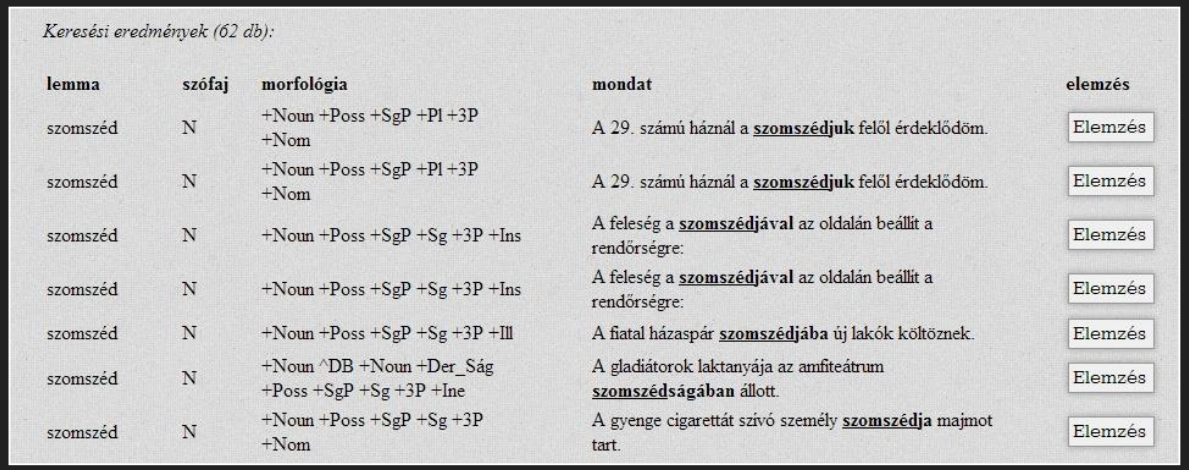
- lemmára vagy szóra történő keresés,
- szűrés morfológiai jegyek és a keresett lemmát/szót tartalmazó összetevők beállítása alapján (szűrés beállítása űrlap segítségével),
- talált elemzések megjelenítése,
- a találati listából kiválasztott mondatelemzés ágrajzának megjelenítése.

http://corpus.hungram.unideb.hu

# 3 Az online keresőfelület (2)

**Találati eredmények megjelenítési formája:**

- eredmények számának megjelenítése;
- az eredménylista oszlopai;
- találatok rendezési elve;
- kiemelési szempontok.



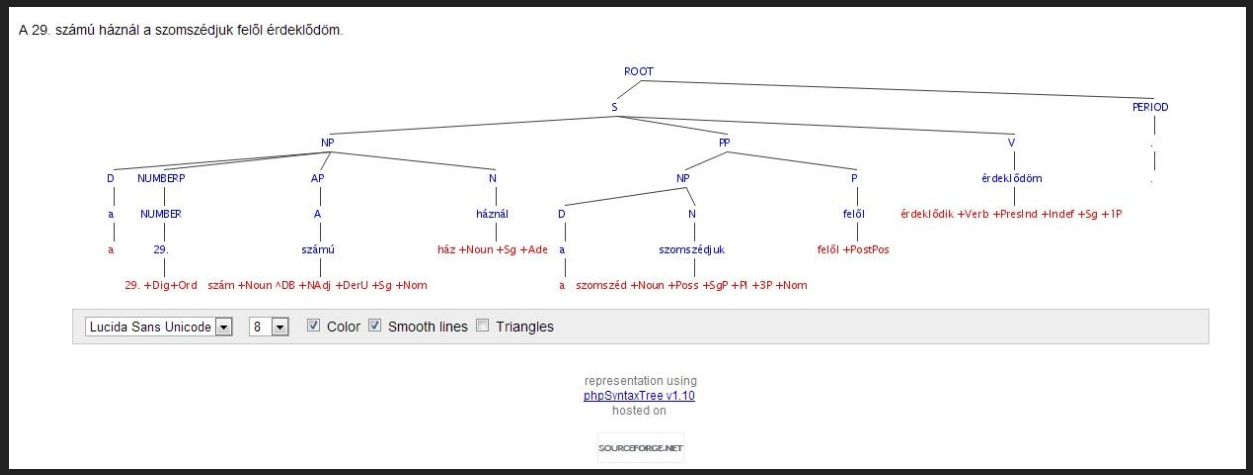
*Keresési eredmények (62 db):*

| lemma | szófaj | morfológia | mondat | elemzés |
|---|---|---|---|---|
| szomszéd | N | +Noun +Poss +SgP +Pl +3P +Nom | A 29. számú háznál a **szomszédjuk** felől érdeklődöm. | Elemzés |
| szomszéd | N | +Noun +Poss +SgP +Pl +3P +Nom | A 29. számú háznál a **szomszédjuk** felől érdeklődöm. | Elemzés |
| szomszéd | N | +Noun +Poss +SgP +Sg +3P +Ins | A feleség a **szomszédjával** az oldalán beállít a rendőrségre: | Elemzés |
| szomszéd | N | +Noun +Poss +SgP +Sg +3P +Ins | A feleség a **szomszédjával** az oldalán beállít a rendőrségre: | Elemzés |
| szomszéd | N | +Noun +Poss +SgP +Sg +3P +Ill | A fiatal házaspár **szomszédjába** új lakók költöznek. | Elemzés |
| szomszéd | N | +Noun ^DB +Noun +Der_Ság +Poss +SgP +Sg +3P +Ine | A gladiátorok laktanyája az amfiteátrum **szomszédságában** állott. | Elemzés |
| szomszéd | N | +Noun +Poss +SgP +Sg +3P +Nom | A gyenge cigarettát szívó személy **szomszédja** majmot tart. | Elemzés |

# 3 Az online keresőfelület (3)

További keresési/megjelenítési tulajdonságok:

- begépeléskor a még lehetséges (és a treebankben ténylegesen előforduló) lemmák/szóalakok listájának megjelenítése;
- talált elemzések vizualizált (ágrajzos) megjelenítése a phpSyntaxTree (v1.10) segítségével.



# Bibliográfia


- **Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., and Varga, D.**: Web-based frequency dictionaries for medium density languages. In: Kilgarriff, A., Baroni, M. (eds.): *Proceedings of the 2nd International Workshop on Web as Corpus ACL-06* (2006) 1–9.
- **Laczkó, T., Rákosi, Gy., Tóth, Á. and Csernyi, G.:** Nyelvtanfejlesztés, implementálás és korpuszépítés: A HunGram 2.0 és a HG-1 Treebank legfontosabb jellemzői. In: Tanács Attila, Vincze Veronika (szerk.): *MSZNY 2013: IX. Magyar Számítógépes Nyelvészeti Konferencia* , 85-96.
- **phpSyntaxTree**
https://code.google.com/p/phpsyntaxtree/
- **The TIGER-XML treebank encoding format**
http://www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/doc/html/TigerXML.html
- **XLE Documentation**
http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html