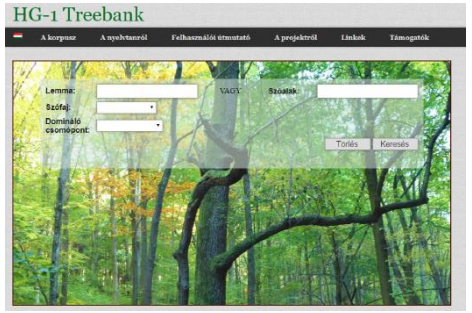


HunGram 2.0 és a HG-2 Treebank – Nyelvtanfejlesztés, implementáció és korpuszépítés

Laczkó Tibor & Rákosi György

Debreceni Egyetem, Angol Nyelvészeti Tanszék

Korpuszok és nyelvtechnológiai eszközök a bölcsészet- és társadalomtudományokban

Debrecen, 2019.02.07.

Az előadás felépítése

- A nyelvelméleti keretünkről
- ParGram
- HunGram
- HunGram-ParGram
- HunGram 2.0
- HG-2 Treebank

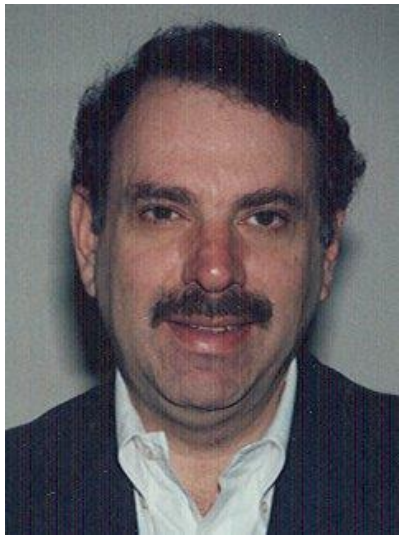
A nyelvelméleti keretünkről (1)

Lexikai-Funkcionális Grammatika (LFG)

- a 70-es évek második fele
- alternatív (nem chomskyánus = nem transzformációs) generatív nyelvtani modell

Joan Bresnan

Stanford University



Ronald Kaplan

Palo Alto Research Center (PARC)

A nyelvelméleti keretünkről (2)

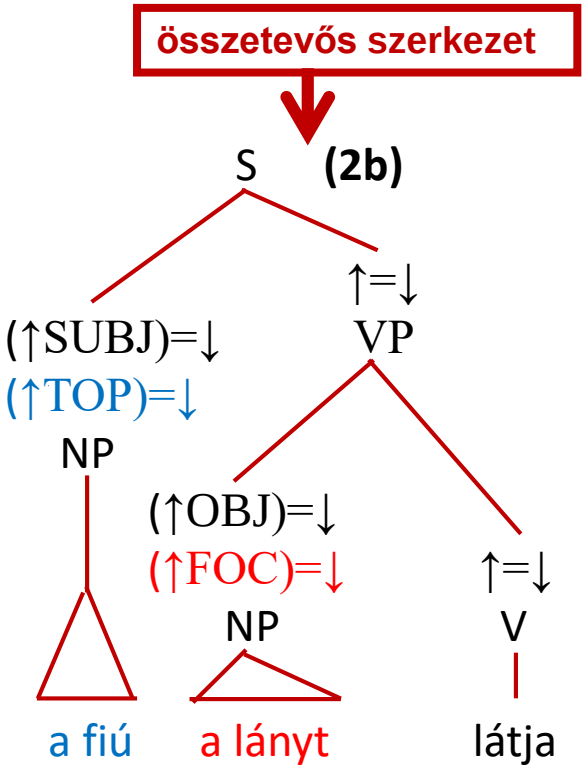
- ***reprezentációs modell***

A nyelvelméleti keretünkről (3)

(2)a. *A fiú látja a lányt.*
b. *A fiú a ’lányt látja.*
c. *Látja a fiú a lányt.*
d. *Látja a lányt a fiú.*
e. *A lányt látja a fiú.*
f. *A lányt a fiú látja.*

Az ö-szerkezetbeli funkciós annotációk biztosítják két szerkezet közötti megfelelő leképezési viszonyokat:

(↑X)=↓ & ↑=↓



funkcionális szerkezet:
függőségi viszonyok, vö. LMNY: „szintaxis”

PRED lát <SUBJ,OBJ>

SUBJ — fiú
TOP — CASE nom

OBJ — lány
FOC — CASE acc

ParGram (1)

- **Par**allel **Gram**mar – *Párhuzamos Nyelvtan*
- Egy LFG alapú, nagyszabású számítógépes nyelvtanfejlesztő projekt, központ: PARC (volt)
- az LFG-t eleve úgy tervezték meg, hogy implementálni lehessen (Ronald Kaplan)
- a közös számítógépes nyelvtanírási platform: XLE (Xerox Linguistic Environment „Xerox Nyelvészeti Környezet” – a PARC és a XEROX cég együttműködéséből)

ParGram (2)

Főbb célkitűzések

- az LFG generatív elméleti modelljének tesztelése: valóban alkalmas keret-e egy egyetemes (univerzális) nyelvtani megközelítés kidolgozására?
  - elvszerűen és pszichológiailag plauzibilis módon megragadhatók-e vele az emberi nyelvek közötti rendszerszerű hasonlóságok és különbözőségek? (vö. funkcionális szerkezet vs. összetevős szerkezet)
- ipari hasznosíthatóság
  - szabály alapú gépi fordítás (a funkcionális szerkezet kiemelt szerepe)
  - intelligens webes keresőprogramok kidolgozása

ParGram (3)

- kiterjedt, aktív, rendszeres nemzetközi együttműködés, számos nyelv beható vizsgálata: angol, német, francia, ír, japán, kínai, norvég, török, vietnámi, arab, spanyol, urdu, indonéz, magyar stb.
- évente egy vagy két műhelytalálkozó
  - **2011: Debrecen** & PARC
  - 2012: Bali
  - **2013: Debrecen** (ParGram Műhelytalálkozó & LFG'13 Nemzetközi Konferencia)
  - 2014: Varsó
  - 2015: Konstanz
- eltökélt törekvés az elemzési eszközök (kategóriák, szabályok stb.) minél egységesebb kialakítására, alkalmazására

HunGram (1)

- 2005/2006, Laczkó Tibor – Fulbright ösztöndíj: Stanford Egyetem & PARC → a HunGram alapvetése

- 2008: Lexikai-Funkcionális Grammatikai Kutatócsoport

HunGram (2)

2008: **Lexikai-Funkcionális Grammatikai Kutatócsoport**
(Debreceni Egyetem, Angol Nyelvészeti Tanszék)

- ***Laczkó Tibor***
  elméleti kutatás, nyelvtanírás, implementálás
- ***Rákosi György***
  elméleti kutatás, nyelvtanírás, implementálás
- ***Tóth Ágoston***
  implementálás, programozás
- ***Csernyi Gábor***
  implementálás, programozás

\+ ***Szűcs Péter***
elméleti kutatás, nyelvtanírás, implementálás

http://hungram.unideb.hu

HunGram (3)

Tóth Ágoston

Laczkó Tibor

Rákosi György

Csernyi Gábor

Szűcs Péter

HunGram (4)

OTKA pályázat, 2008-2013

- a magyar nyelv lexikai-funkcionális grammatikája
- a grammatika implementálása az XLE platformján ParGram-es könyezetben: HunGram 1.0

TÁMOP (kutatóegyetemi) pályázat, 2010-2012

- HunGram 2.0 kifejlesztése
- HG-2 Treebank létrehozása (a HunGram 2.0 révén)

HunGram (5)

TOKENIZÁLÓ	⇨	MORFOLÓGIAI ELEMZŐ	⇨	LEXIKON	⇨	SZINTAXIS
tokenizálás		morfológiai elemzés		lexikai kikeresés		szerkezeti elemzés

HunGram (6)

- a **tokenizáló** az adott szófűzért tokenekre bontja

→bemenet a **morfológiai komponens**, egy véges állapotú átalakító (*finite state transducer* (**FST**)) számára

- az FST tageket rendel minden morfhoz

  (1) *játékot:* játék "+Noun" "+Sg" "+Acc"

  (2) *ették:* eszik "+Verb" "+Past" "+Def" "+Pl" "+3P"

- a tageknek (funkcionális annotációval is ellátott) önálló lexikai tételeik vannak

  (3) +Acc N_SFX XLE (^ CASE)= acc.

  (4) +Pl N_SFX XLE (^ NUM)= pl;

  V_SFX XLE (^ SUBJ NUM)= pl

HunGram (7)

hungram1.lfg

```
FIRST   HUNGARIAN   CONFIG (1.0)
ROOTCAT   ROOT.
FILES common.templates.lfg hun_lex_core.lfg hun_lex_p.lfg hun_templates.lfg
hun_morphconfig.lfg hun_rules.lfg common.features.lfg hungarian.features.lfg.
LEXENTRIES  (CORE HUNGARIAN) (PARTICLES_AND_POSTPOSITIONS
HUNGARIAN).
CHARACTERENCODING utf-8.
MORPHOLOGY (STANDARD HUNGARIAN).
RULES   (FIRST HUNGARIAN).
FEATURES      (STANDARD COMMON) (FIRST HUNGARIAN).
TEMPLATES   (STANDARD COMMON) (FIRST HUNGARIAN).
GOVERNABLERELATIONS   SUBJ OBJ POSS OBL OBL-?+ COMP XCOMP PREDLINK.
```

hun_rules.lfg

```
FIRST   HUNGARIAN   RULES (1.0)

"*****************************
NOUN PHRASE RULES
*****************************"
```

```
DP -->  {(DP: { (^ POSS)=!
                (! CASE)=c dat
                (^ CHECK _POSS-MORPH)=c +
              | (^ SPEC)= !
                (! NUM)=(^ NUM)
                (! CASE)=(^ CASE)
                (! PRON-TYPE)=c demon } )
          D'
        | PRON }.
```

a) **Jánosnak** a könyvét
b) **ezt** a köyvet

HunGram (8)

hun_lex_core.lfg

CORE HUNGARIAN LEXICON (1.0)

"**********VERBS
**********"

```
hív     V XLE {(^PRED)='hív <(^SUBJ) (^OBJ)>'
               | @(CONCAT (^ PRT-FORM) `# %stem %FN)
               (^ PRED)='%FN <(^SUBJ) (^OBJ)>'
               (^ CHECK _PRT-VERB)=+
               {(^ PRT-FORM) =c el | (^ PRT-FORM) =c fel}}.
```

"**********
TAGS
**********"

```
+Noun     N_SFX XLE  @(NSYN common)
+Acc      N_SFX XLE @(UP-CASE acc)
```

hun_templates.lfg

FIRST HUNGARIAN TEMPLATES (1.0)

ADJUNCT = ! $ (^ ADJUNCT). "assigns the ADJUNCT GF to non-DPs (as member of set)"

@ADJUNCT

HunGram (9)

hun_morphconfig.lfg

STANDARD HUNGARIAN MORPHOLOGY (1.0)

TOKENIZE:
P!initdown.fst prts.fst basic-parse-tok.fst
#G!default-gen-tok.fst

ANALYZE:
hun-lexicon.fst rm.fst

hungarian.features.lfg

FIRST HUNGARIAN FEATURES (1.0)

PRT-FORM: -> $ { agyon át be bele benn benne bent el érte észre fel föl hozzá ide keresztül ki le meg neki rá rajta oda össze szét vissza volna utána}.
+CASE: -> $ { inessive elative superessive essive-formal terminative ablative allative adessive illative delative sublative translative causal-final }.

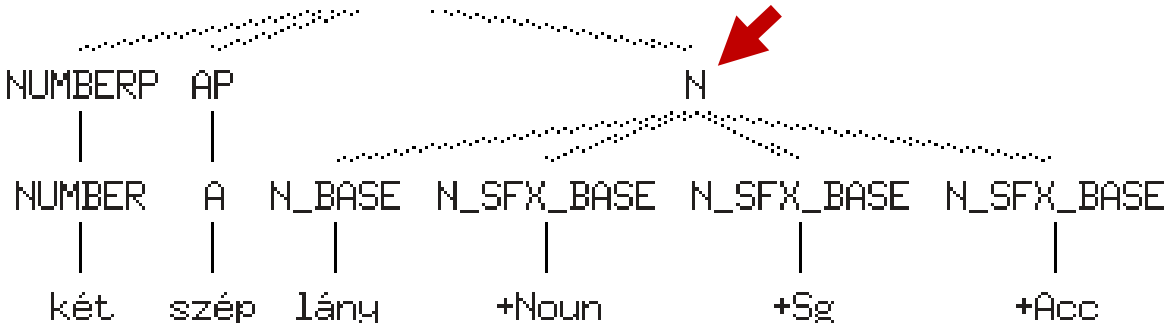
HunGram (10)

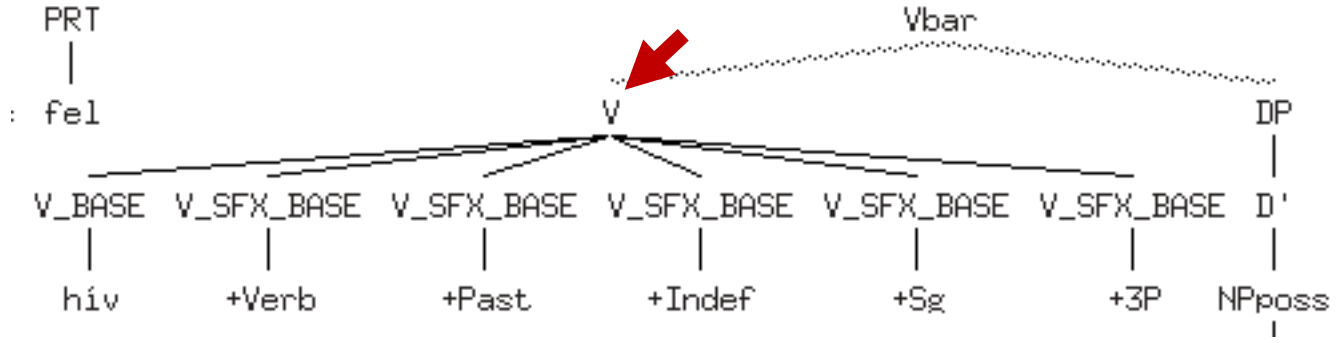
szublexikális (= szószint alatti) szerkezet, funkcionális információ

	SZÓ	TAG-SPECIFIKÁCIÓ
a.	hív-ott	hív “+Verb” “+Past” “+Indef” “+Sg” “3P”
b.	lány-t	lány “+Noun” “+Sg” “+Acc”

```
N -->   N_BASE
        N_SFX_BASE*.
```

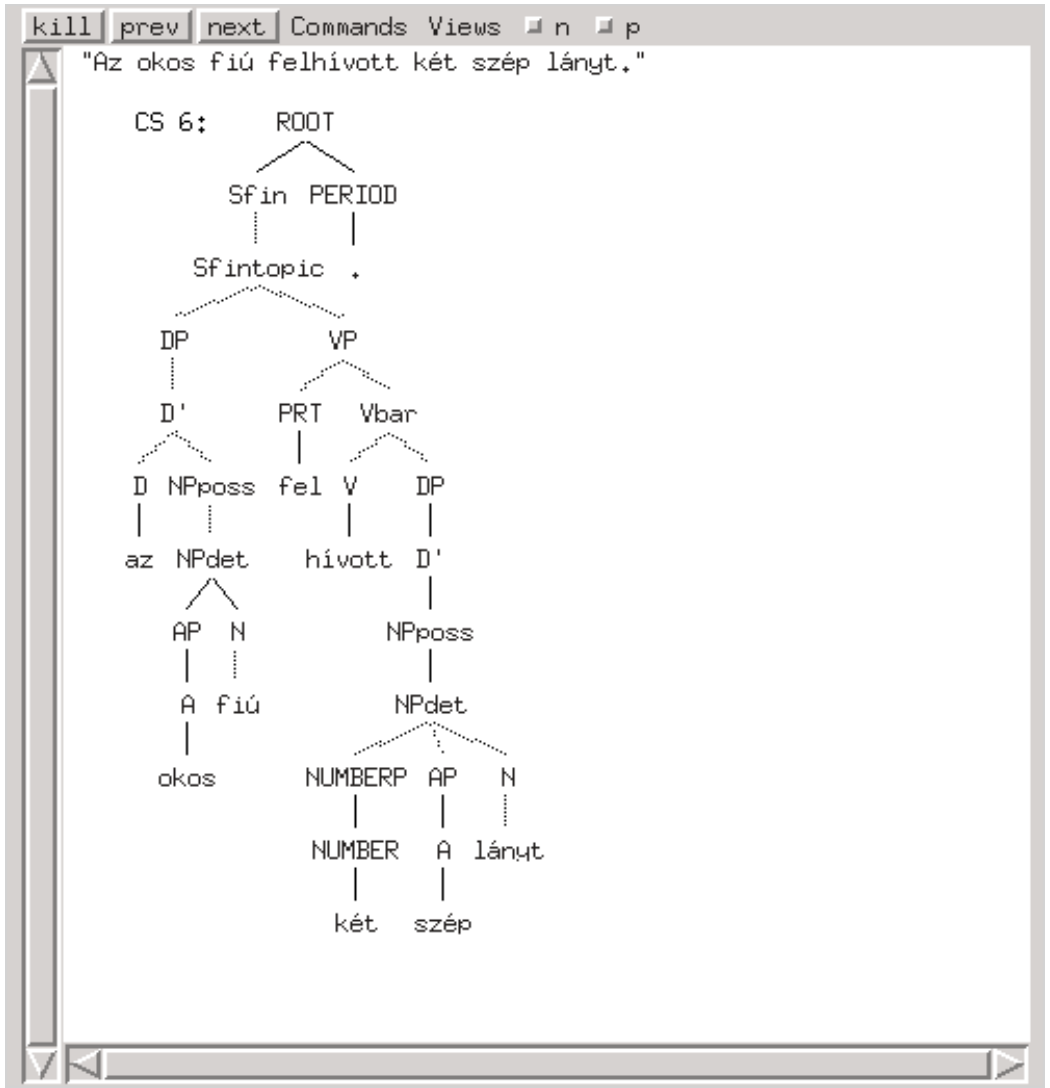
HunGram (11)

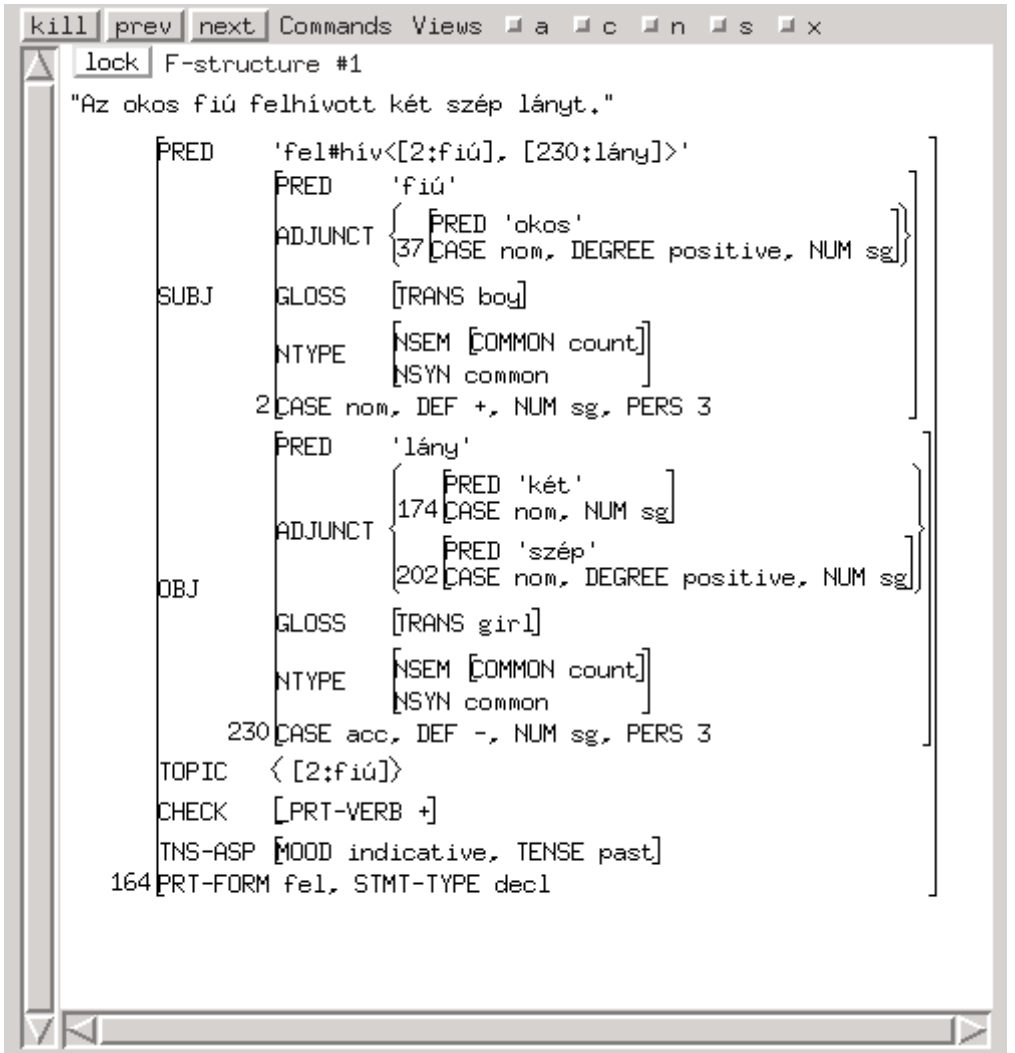

NUMBERP
AP
N
NUMBER
A
N_BASE
N_SFX_BASE
N_SFX_BASE
N_SFX_BASE
két
szép
lány
+Noun
+Sg
+Acc


PRT
Vbar
fel
V
DP
V_BASE
V_SFX_BASE
V_SFX_BASE
V_SFX_BASE
V_SFX_BASE
V_SFX_BASE
D'
hív
+Verb
+Past
+Indef
+Sg
+3P
NPposs

HunGram (12)

Az okos fiú felhívott két szép lányt.





HunGram-ParGram (1)

Links to ParGram Groups

Here are some links to the Wikis or sites of individual grammar groups. It might be useful to check out languages that are similar to the one you wish to work on (or are working on). If you would like to obtain a particular ParGram grammar, you should contact the groups directly. For example, the Polish grammar is available under the GNU General Public License (version 3).

HunGram Hungarian Grammar: http://hungram.unideb.hu

POLFIE Polish LFG/XLE grammar: http://zil.ipipan.waw.pl/LFG

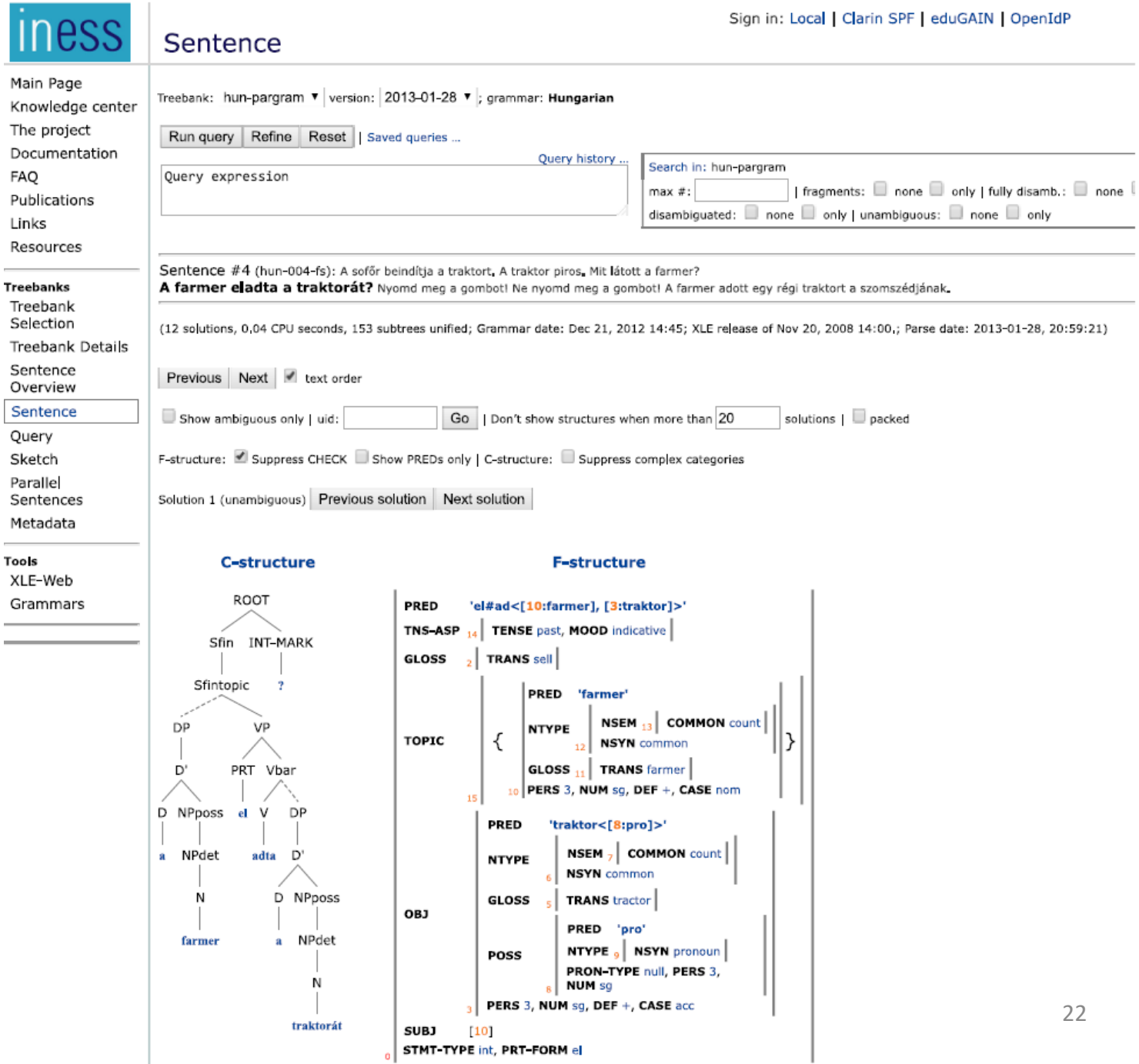
Urdu ParGram Grammar: http://ling.uni-konstanz.de/pages/home/pargram_urdu

Norwegian ParGram Grammar: http://clarino.uib.no/redmine/projects/inesspublic/wiki/NorGram_documentation

Sulger, Sebastian, Miriam Butt, Tracy Holloway King, Paul Meurer, **Tibor Laczkó, György Rákosi,** Cheikh Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, Agnieszka Patejuk, Ozlem Cetinoglu, I Wayan Arka and Meladel Mistica. 2013.ParGramBank: The ParGram Parallel Treebank. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 4-9, 2013, 550–560. (14 társszerző)

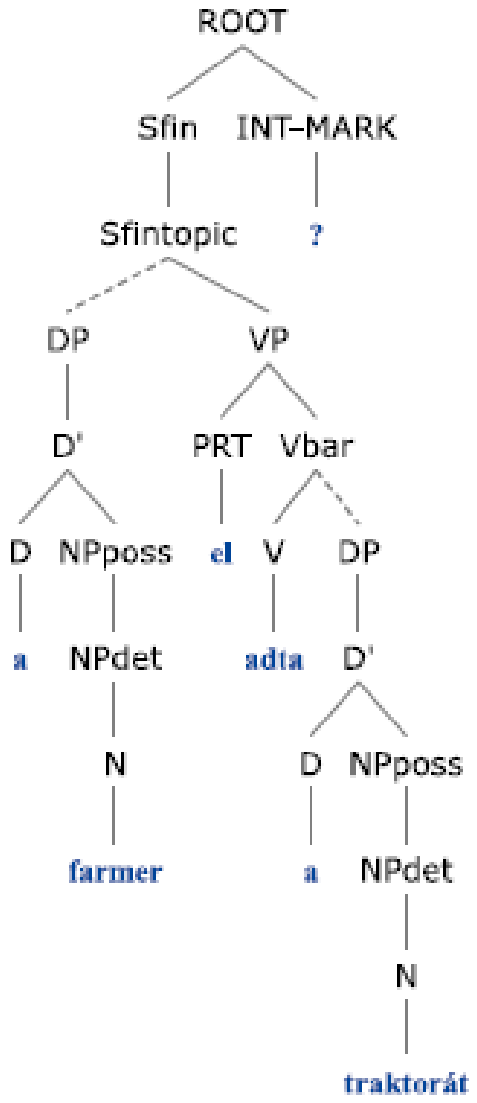
HunGram-ParGram (2)

Parallel treebankek



HunGram-ParGram (3)

C-structure



F-structure

```
PRED      'el#ad<[10:farmer], [3:traktor]>'
TNS-ASP   14 | TENSE past, MOOD indicative |
GLOSS     2  | TRANS sell |
TOPIC     15 { PRED   'farmer'
               NTYPE  12 | NSEM 13 | COMMON count |
                         | NSYN common |
               GLOSS  11 | TRANS farmer |
               10 | PERS 3, NUM sg, DEF +, CASE nom | }
OBJ       3  | PRED   'traktor<[8:pro]>'
             | NTYPE  6 | NSEM 7 | COMMON count |
             |        | NSYN common |
             | GLOSS  5 | TRANS tractor |
             | POSS   8 | PRED   'pro'
             |          | NTYPE 9 | NSYN pronoun |
             |          | PRON-TYPE null, PERS 3,
             |          | NUM sg |
             | PERS 3, NUM sg, DEF +, CASE acc |
SUBJ      [10]
0 STMT-TYPE int, PRT-FORM el
```

HunGram 2.0 (1)

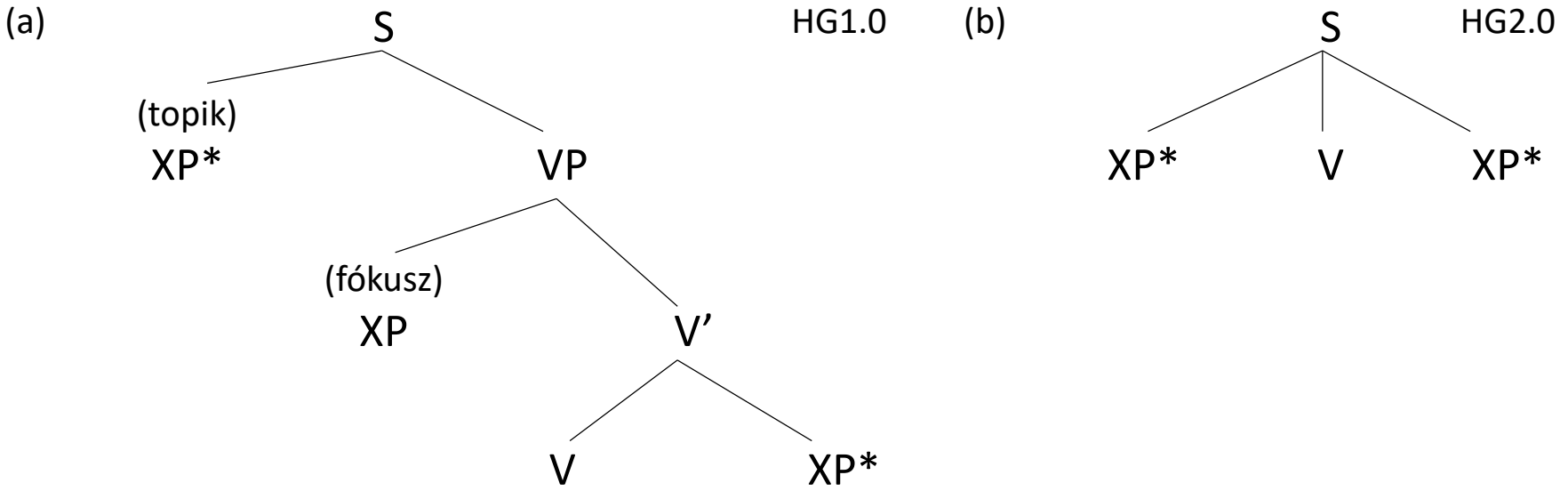
cél: a treebank kielégítése

- automatikus elemzés
- minél kevesebb többértelműség
- DE: nyelvészeti szempontból teljes és megbízható elemzés, nem „sekély(es)” nyelvtan
- inkább: pontosság, kevésbé: lefedettség
- menetközben volt (több körös) (véletlenszerű) tesztelés, módosítás, javítás → kiérlelt nyelvtanváltozat
- csak ö-szerkezet, nincs f-szerkezet

HunGram 2.0 (2)

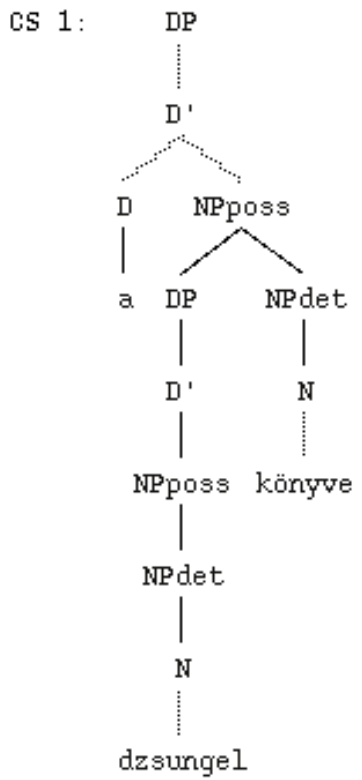
leegyszerűsített (jóval laposabb) frázisstruktúrák

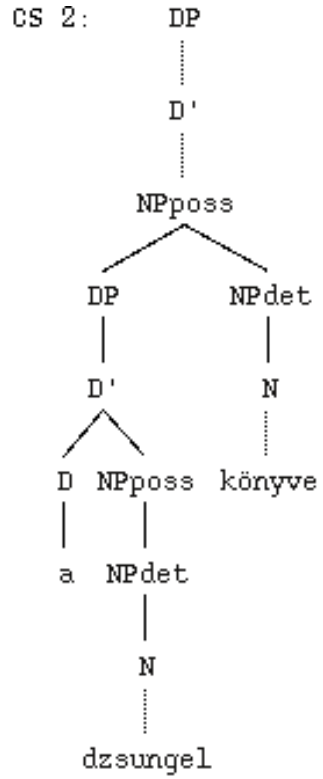
- kevésbé hierarchikus főnévi csoport
- lapos mondatszerkezet (pl. nincs kitüntetett topik és fókusz pozíció)

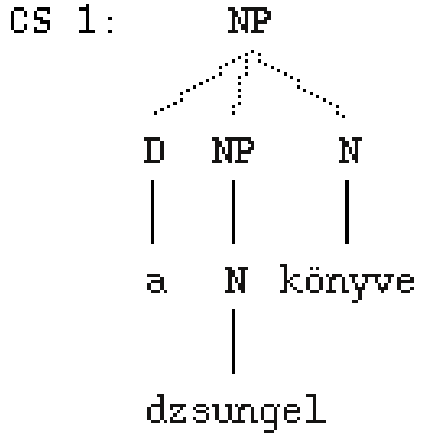


HunGram 2.0 (3)

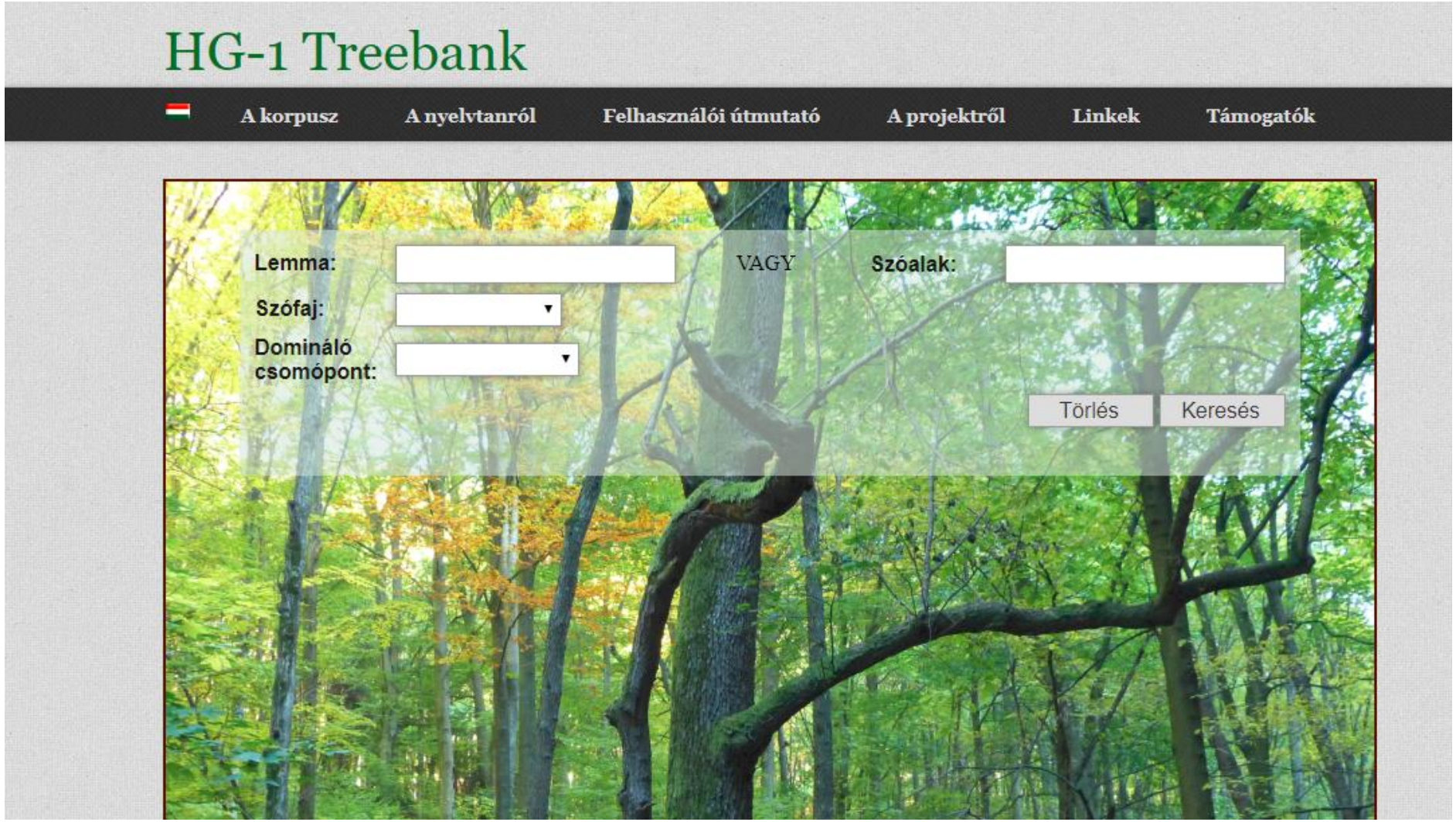
kizártuk a pusztán nyelvészeti szempontból érdekes kétértelműségeket: *a dzsungel könyve*







27 Treebank


HG-1 Treebank
A korpusz
A nyelvtanról
Felhasználói útmutató
A projektről
Linkek
Támogatók
Lemma:
VAGY
Szóalak:
Szófaj:
Domináló csomópont:
Törlés
Keresés

28 Treebank: Jellemzők

- HunGram 2.0 nyelvtanra épülő (első) treebank
- tagsetek (XLE)
  - morfológia: XEROX - Basis Technology
  - mondattan: HunGram 2.0

- forrás: Magyar Webkorpusz
- 1.5 millió szó
  - 1 elemzés: 152778 (752159 szó)
  - 2 elemzés: 131008 (771716 szó)

29 Treebank: Implementációs feladatok

- Mondatok elemeztetése a HunGram nyelvtannal feltöltött XLE elemzővel, és a kimenet rögzítése (alternatív elemzésekkel).
- Az összes lehetséges elemzés összetevős szerkezetének kibontása és tárolása. A korpuszt ettől a ponttól XML dokumentumban tároljuk (TigerXML formátumban).
- Alkorpuszok kezelése:
  - korpuszfájlok darabolása és egyesítése,
  - statisztikák készítése (faszélesség, famélység, szavak és mondatok száma, elemzések száma)
- MySQL back-end adatbázis
- On-line lekérdezési felület, ahol szavakra vagy lemmákra kereshetünk, ehhez szűréseket állíthatunk be.

30 Treebank: Online lekérdező felület

- keresés szóra vagy lemmára
- keresés szűrése morfológiai jegyekre és a keresett szót tartalmazó összetevőre (szűrés beállítása űrlap segítségével)
- a találatok megjelenítése
- a találati listából kiválasztott mondatelemzés ágrajzának megjelenítése.

31 Treebank: Online lekérdező felület

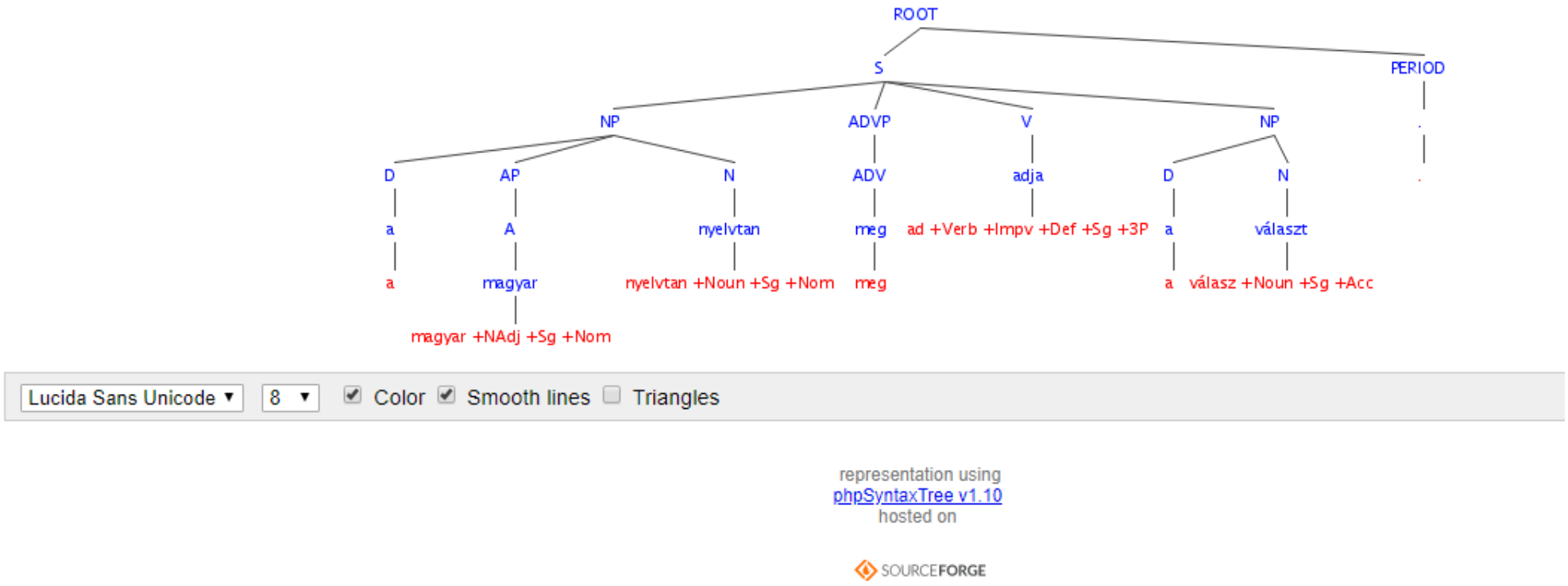
- keresési eredmények számának megjelenítése
- találati eredmények oszlopai
- találatok rendezési elve
- kiemelési szempontok

Keresési eredmények (14 db) a(z) **nyelvtan** *szóalakra:*

lemma	morfológia	mondat	elemzés
nyelvtan	+Noun +Sg +Nom	A konstrukciós **nyelvtan** az amerikai strukturalizmus nyelvészeti hagyományait eleveníti fel.	Elemzés
nyelvtan	+Noun +Sg +Nom	A konstrukciós **nyelvtan** az amerikai strukturalizmus nyelvészeti hagyományait eleveníti fel.	Elemzés
nyelvtan	+Noun +Sg +Nom	A lemez a magyar **nyelvtan** szabályait tartalmazza.	Elemzés
nyelvtan	+Noun +Sg +Nom	A lemez a magyar **nyelvtan** szabályait tartalmazza.	Elemzés

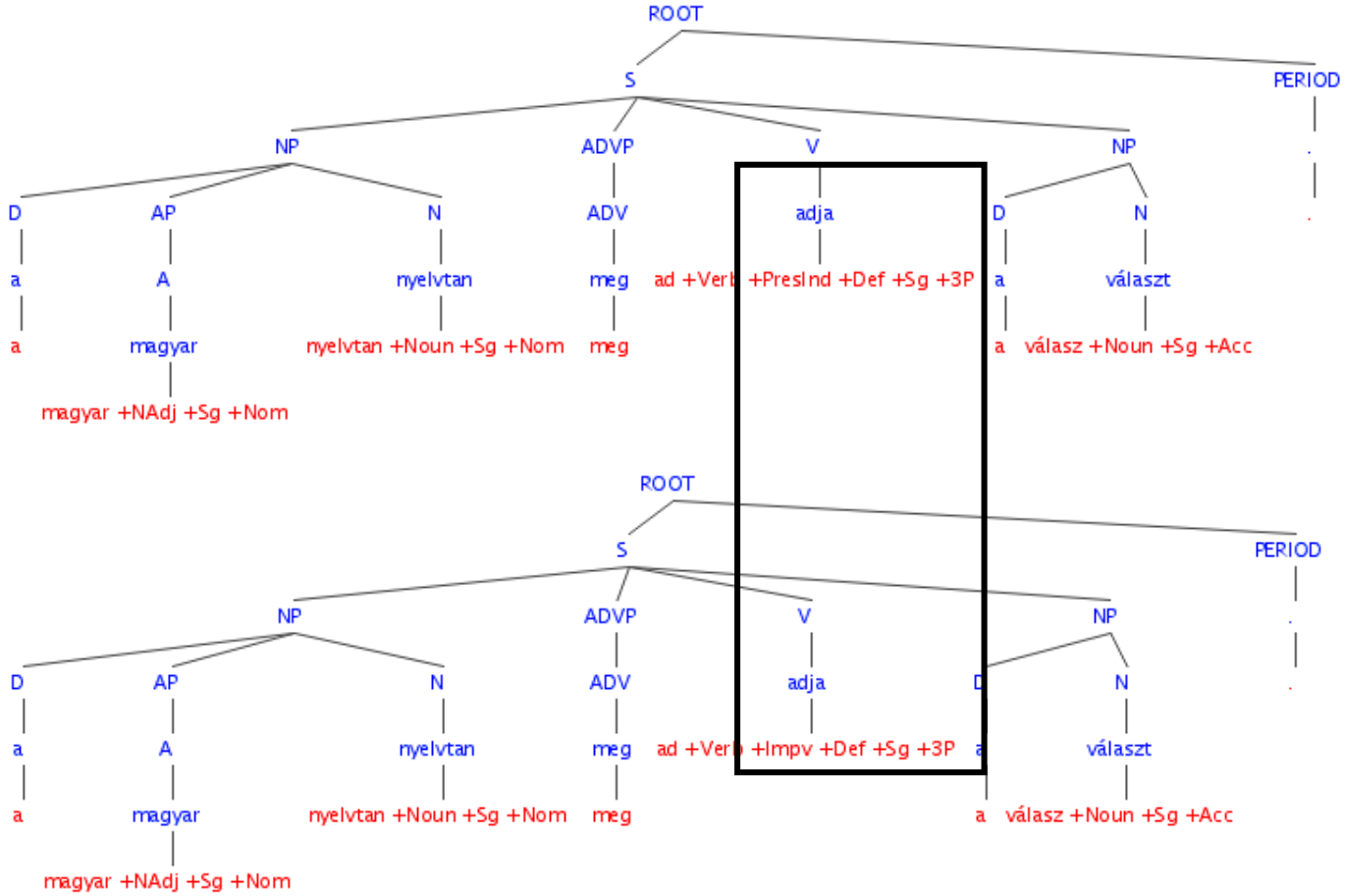
32 Treebank: Online lekérdező felület

- talált elemzések megjelenítése a phpSyntaxTree (v1.10) segítségével.



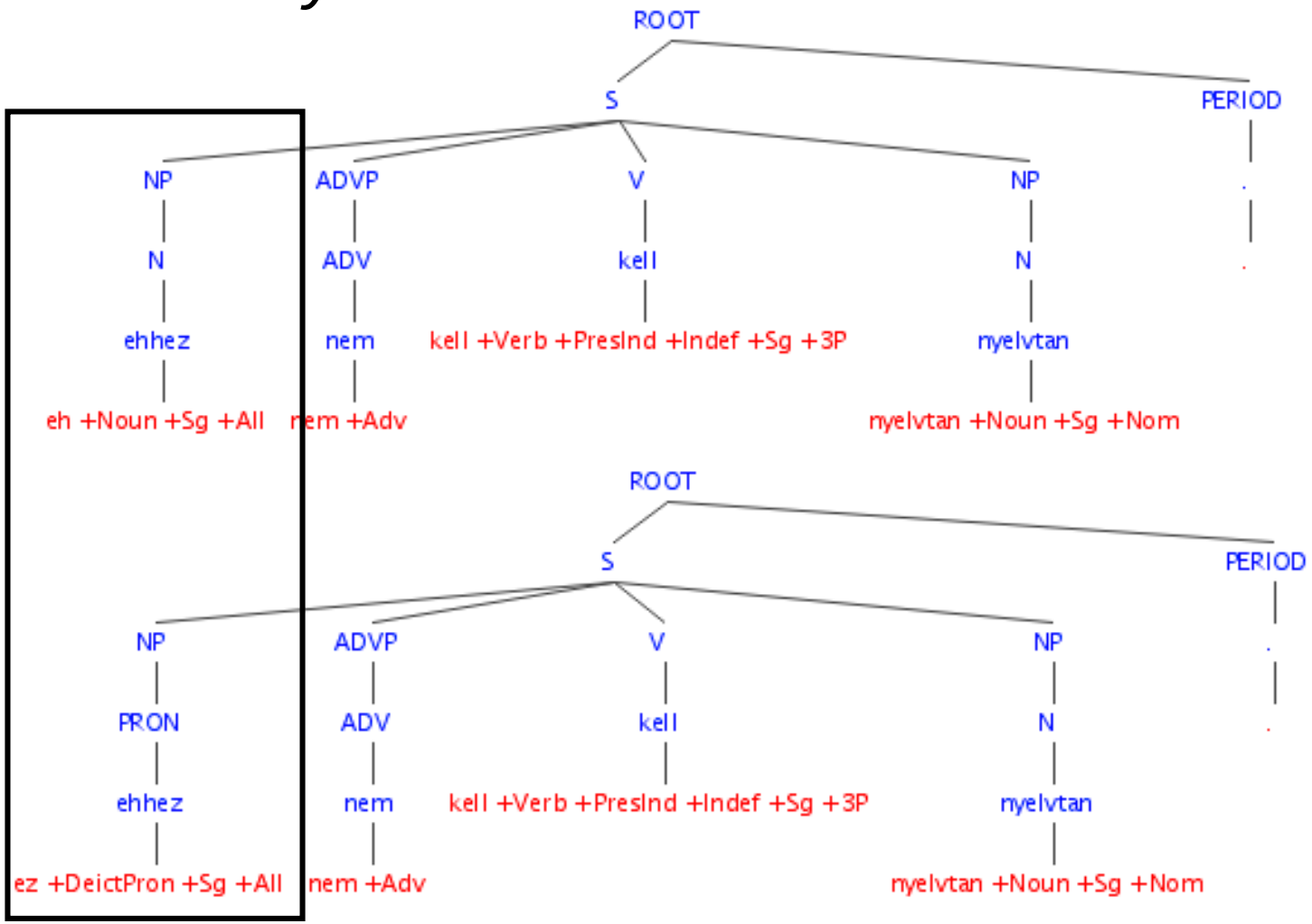
33 Treebank: Kétértelműségek

A magyar nyelvtan megadja a választ.



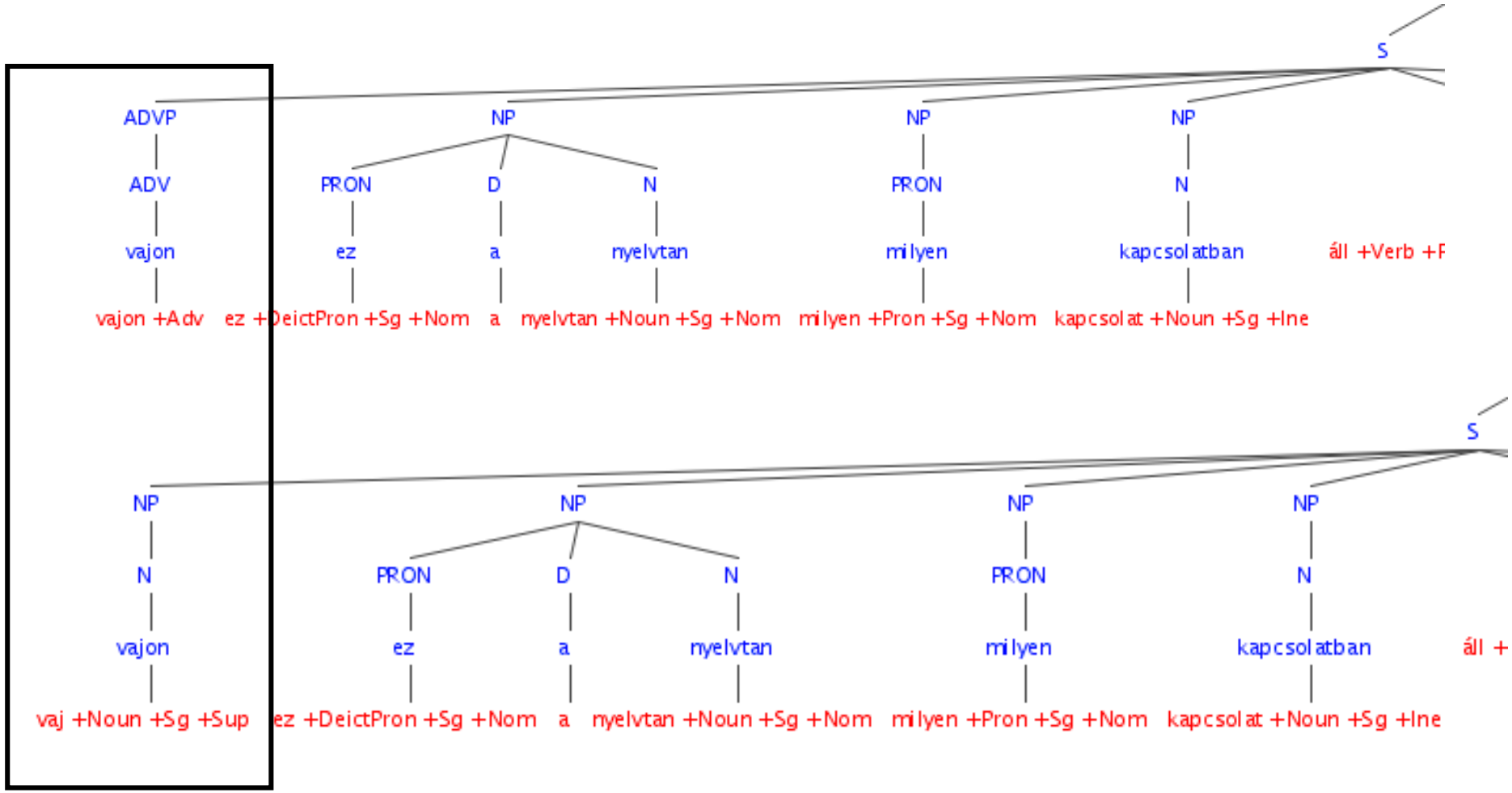
34 Treebank: Kétértelműségek

Ehhez nem kell nyelvtan.



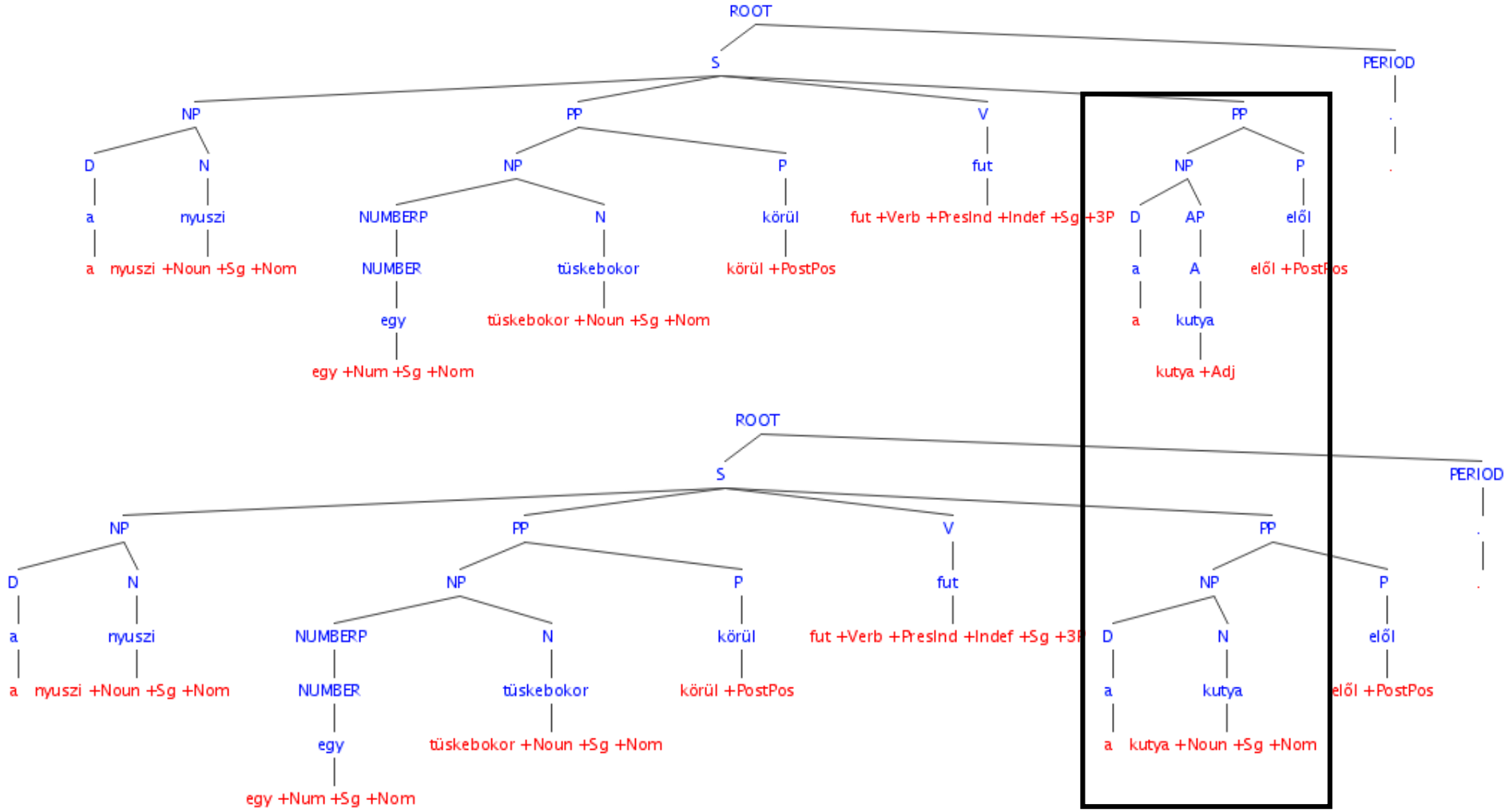
35 Treebank: Kétértelműségek

Vajon ez milyen kapcsolatban áll az eleven nyelv grammatikájával?



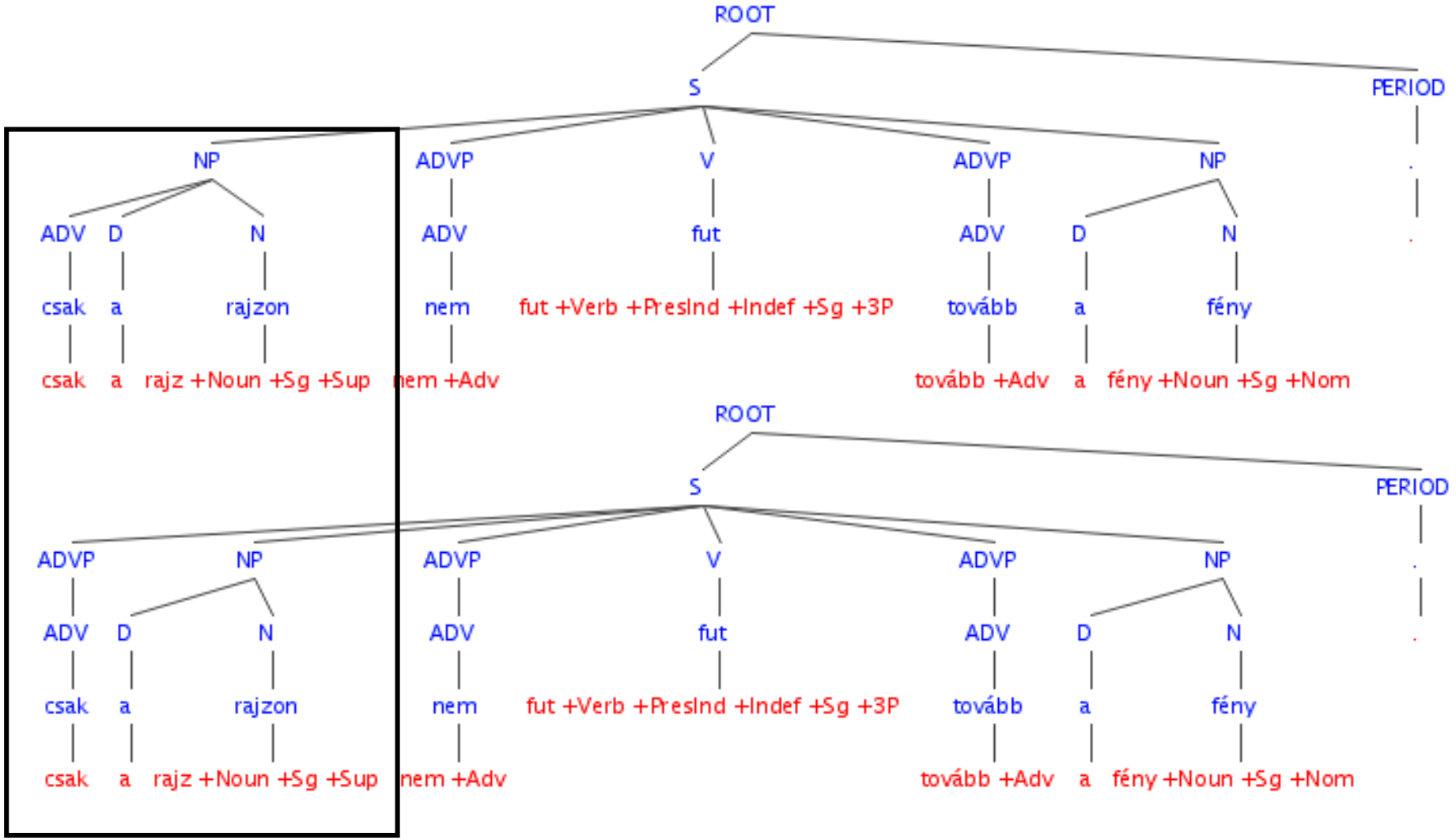
36 Treebank: Kétértelműségek

A nyuszi egy tüskebokor körül fut a kutya elől.



37 Treebank: Kétértelműségek

Csak a rajzon nem fut tovább a fény.



38 Treebank: A projekt tagjai

Laczkó Tibor

- A HunGram Kutatócsoport vezetője
- nyelvtanírás, nyelvészeti háttérkutatások

Rákosi György

- nyelvtanírás, nyelvészeti háttérkutatások

Tóth Ágoston

- korpuszmondatok elemzése (XLE elemző, NIIF szuperszámítógépes szolgáltatás felhasználásával)
- a korpuszt tartalmazó Tiger-XML-alapú leírónyelv elemeinek kidolgozása

Csernyi Gábor

- adatbázistervezés és implementálás a webes interfészhez
- XLE kimenet átalakításának programozási munkái
- webes interfész programozása

39 Közvetlen tervek: Egy névmási korpusz

- A jelenleg futó OTKA-projektünkben végzett kutatások részeként a HunGram 2.0 fejlesztése a névmásokra és a névmásokat tartalmazó szerkezetekre fókuszálva.

- Ezzel a bővített nyelvtanváltozattal egy újabb, névmási elemeket tartalmazó mondatokra specializált korpusz létrehozása.

- A nyelvtanszfejlesztési munkákba Szűcs Péter is bekapcsolódott.

40 Köszönetnyilvánítás, referencia

A 111918 számú projekt (*Új megközelítések a magyar névmások nyelvtanának leírásában*) a Nemzeti Kutatási Fejlesztési és Innovációs Alapból biztosított támogatással, az K pályázati program finanszírozásában valósult meg.

Laczkó Tibor, Rákosi György, Tóth Ágoston & Csernyi Gábor. 2013. Nyelvtanfejlesztés, implementálás és korpuszépítés: A HunGram 2.0 és a HG-1 Treebank legfontosabb jellemzői. In Tanács Attila & Vincze Veronika szerk. *IX. Magyar Számítógépes Nyelvészeti Konferencia*. Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport. 85-96.