Nyelvtechnológia
3

BME, 2007. november 20.

Prószéky Gábor

Morfo-fonológiai „guesser”

kacsónak + 0	✓	**FN** + 0
kacsóna + k	*	FN + PL
kacsón + ak	*	FN + PL
kacsó + nak	✓	**FN** + DAT
kacsó + nak	*	IGE + PL3
kacs + ó + nak	*	IGE + MNI + DAT
ka | csónak + 0	*	FN|FN

Tipikus hibák a számítógéppel létrehozott dokumentumokban

- karakterhibák
- valódi helyesírási hibák
- nyelvhelyességi hibák
- tipográfiai hibák
- helyesírás-ellenőrzés a szavak szintjén
- a szóellenőrzés és a nyelvhelyesség-ellenőrzés viszonya
- a nyelvi programrendszer lehetséges hibái (*kör/kőr*, *-ít*)

A szóellenőrzés menete

(1) Morfológiai elemzés

kérdésse ☞ <nincs ilyen szó a magyarban>

(2) Ajánlás

törlés:

érdésse, krdésse, kérésse, kédésse, kérdése, kérdéss

helycsere:

ékrdésse, krédésse, kérédsse, …, kérdéses

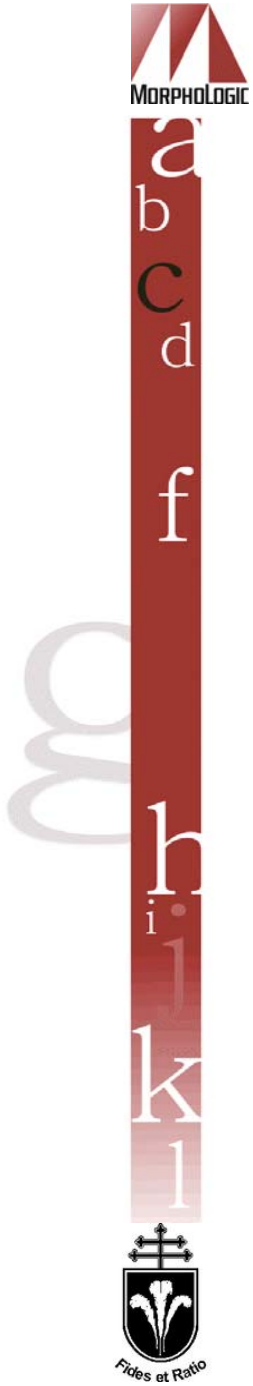
nyelvspecifikus csere:

kérdéssé, kérdesse, …

…

(3) Ellenőrzés morfológiai elemzéssel

kérdése, kérdéses, kérdesse, kérdéssé



Szóellenőrzés morfológiával

kérdése	
kérdés[FN]+e[PSe3]	főnévi
kérd[IGE]+és[IF]+e[PSe3]	főnévi
kérdéses	
kérdéses[MN]	melléknévi
kérdés[FN]+es[SKEP]	melléknévi
kérd[IGE]+és[IF]+es[SKEP]	melléknévi
kérdesse	
kérd[IGE]+es[MUV]+se[TPe3]	igei
kérdéssé	
kérdés[FN]+sé[FAC]	főnévi
kérd[IGE]+és[IF]+sé[FAC]	főnévi

MorphoLogic

Fides et Ratio

Nyelvhelyesség-ellenőrzés a szóhatáron túl

- lehetséges-e mondatszintű helyesírás-ellenőrzés?
- „grammar checker” ?
- parciális elemzések
- hiba-nyelvtan vs. nyelvtan
- hibaelemzések, a hibák súlyozása
- a hiba és a nem-hiba határának elmosódása
- a nyelvi vagy a formai természetű hibák szűrésének preferálása
- stílusellenőrzés számítógéppel

A magyar elválasztás szabályai

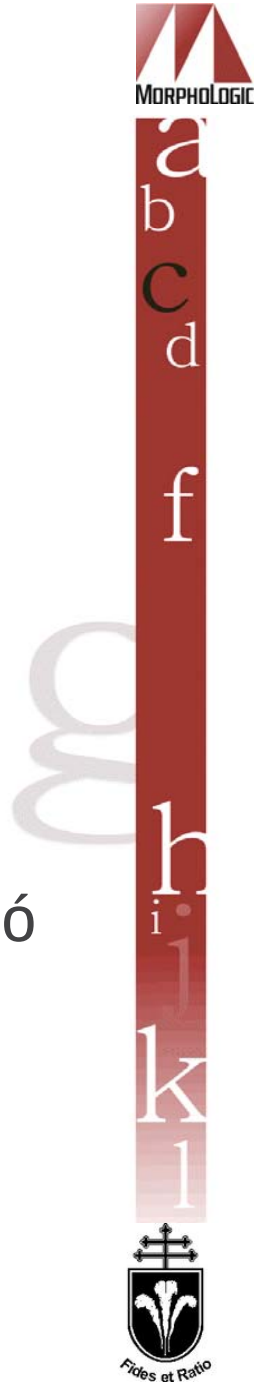
Alap	Elválasztva	Példa
VV	V–V	*ba-uxit*
VC_1C_2V	$VC_1–C_2V$	*er-kély*
VC_iC_iV	$VC_i–C_iV$	*vet-tem*
VCc_1c_2V	$VC–c_1c_2V$	*mor-zsa*
Vc_1c_2CV	$Vc_1c_2–CV$	*asz-tal*
$Vc_{11}c_{12}c_{21}c_{22}V$	$Vc_{11}c_{12}–c_{21}c_{22}V$	*tarisz-nya*
$Vc_1c_1c_2V$	$Vc_1\mathbf{c_2}–c_1c_2V$	*össze/ösz-sze*
#VV	#VV	*autó*
#VC	#VC	*alaki*
VV#	VV#	*hazai*

Automatikus szövegelválasztás

- az elválasztás alkalmazása
- automatikus és interaktív módszerek
- a morfológiai felülbírálás kérdése
- alternatív elválasztások kezelése (többértelműség, illetve a szabályok „engedékenysége” miatt)
- tipográfiai szempontok
- különleges elválasztások (hosszú kettős mássalhangzók, mássalhangzó-háromszorozódás) helyes kezelése

Számítógépes szinonimaszótárak és tezauruszok

- a szinonimákról
- szinonimaszótár vagy tezaurusz?
- tárolási és keresési problémák
- a rokonértelműség definíciója
- az automatikus csere problémái
- tővisszaállítás
- többértelműségek kezelése
- a lexikai és a szintaktikai szó különbségéből adódó nehézségek
- az összetett szavak szinonimáinak problémája
- morfológiai generálás minta alapján

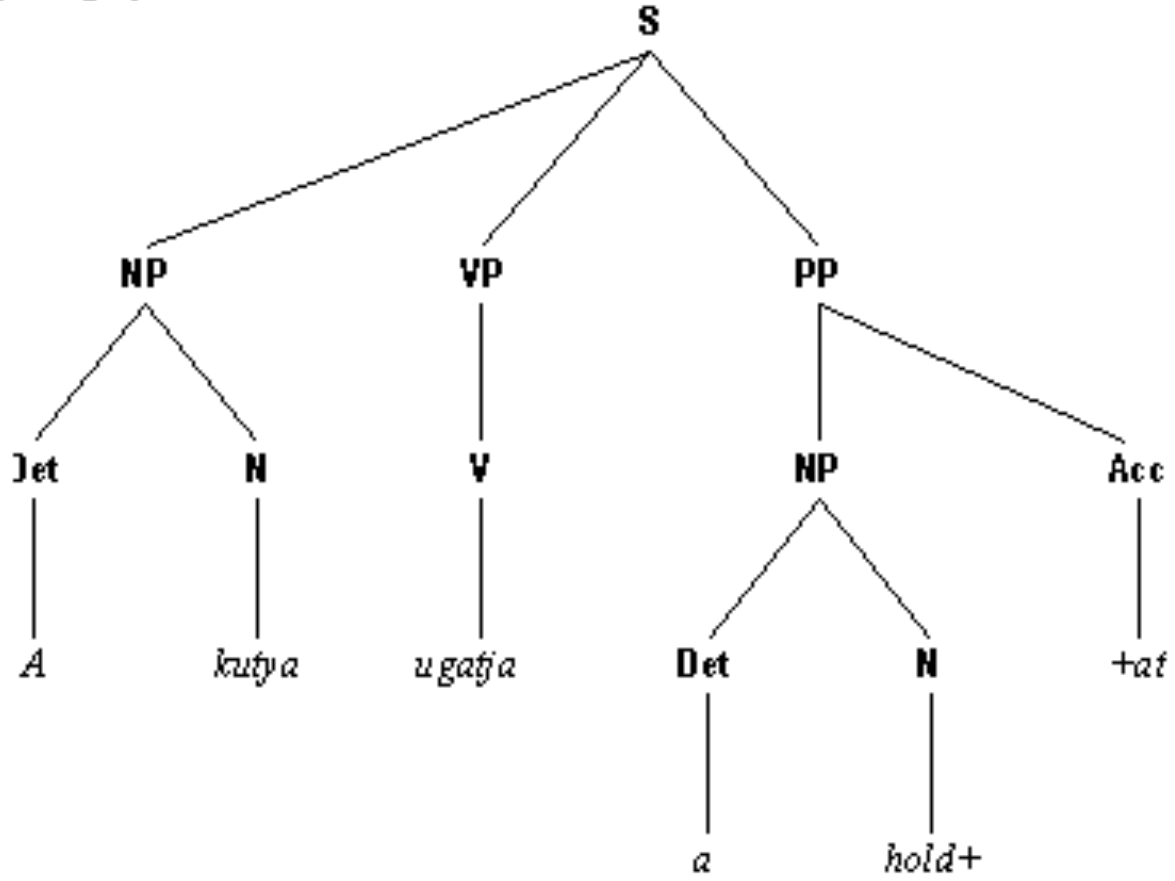


Szintaxis

- közvetlen összetevős szerkezet
- függőségi szerkezet

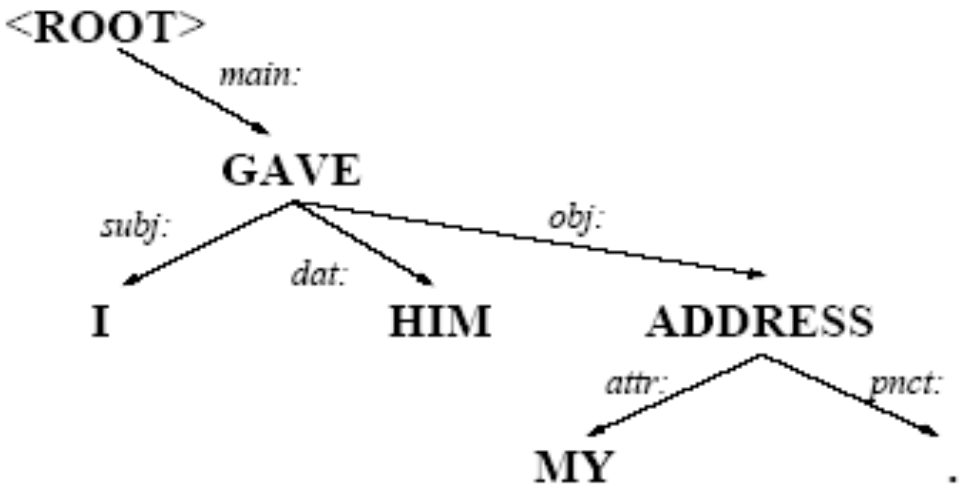
Összetevős szerkezet

A kutya ugatja a holdat.



Függőségi szerkezet

I gave him my address.



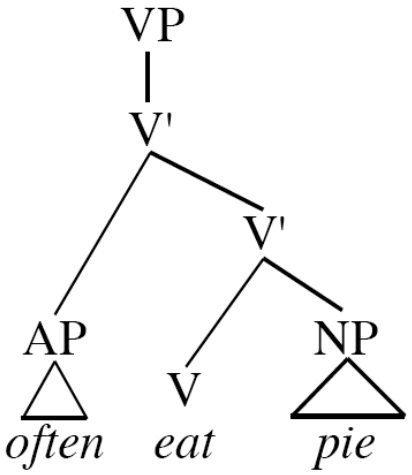
A mondatszerkezet leírásának főbb eszközei

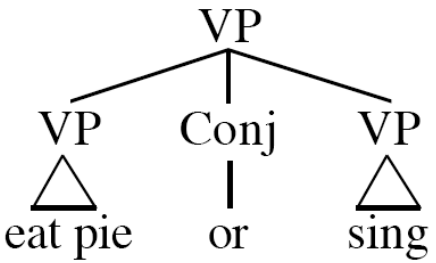
- Közvetlen összetevős nyelvtanok: előnyük a magasabb szintű kategóriák bevezetésének lehetősége, hátrányuk a szintaktikai viszonyok egy részének „kifejezhetetlensége”
- Függőségi szerkezet: előnyük a szintaktikai függőség kifejezésének lehetősége, hátrányuk a magasabb szintű kategóriák kezelhetetlensége
- Egy elegáns közös megoldás: az X-vonás nyelvtanok

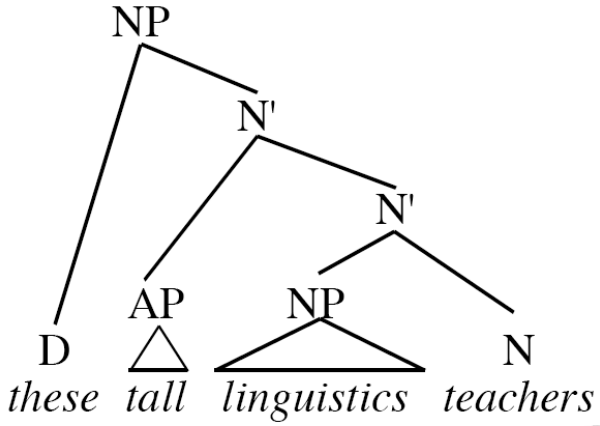
X-vonás: összetevők és függőség

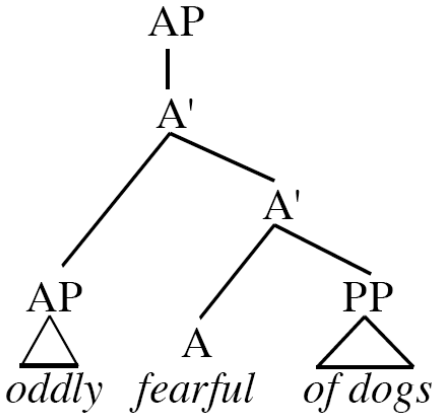
- S → NP VP
- Az összetevős szerkezetben az NP és a VP „testvérek”, azaz mindketten az S „gyermekei”, de ezt nem fejezi ki a függőségi leírás
- Azt viszont a közvetlen összetevős leírás nem fejezi ki, hogy testvérek bár, de nem egyforma súllyal, ui. a VP a szerkezet feje
- X-vonás szabályként: V” → N’ V’
- Azaz: a V” a V maximális projekciója, tehát a mondat feje az ige!
- Csak endocentrikus szerkezetekre! (v.ö. exocentrikus)

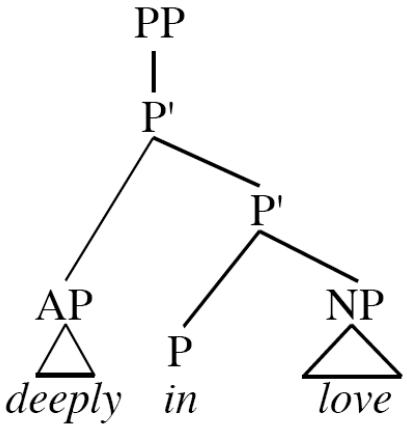
X-vonás szerkezetek



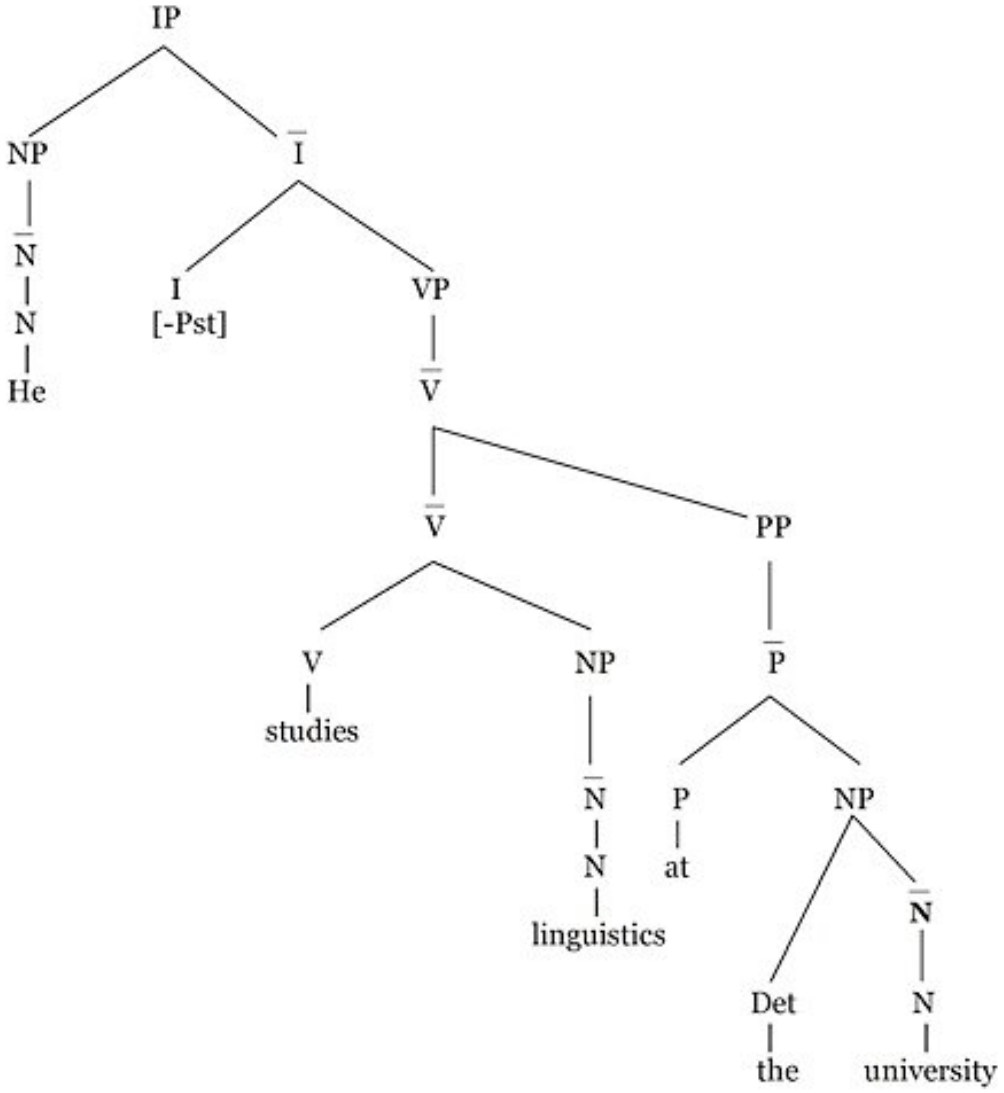




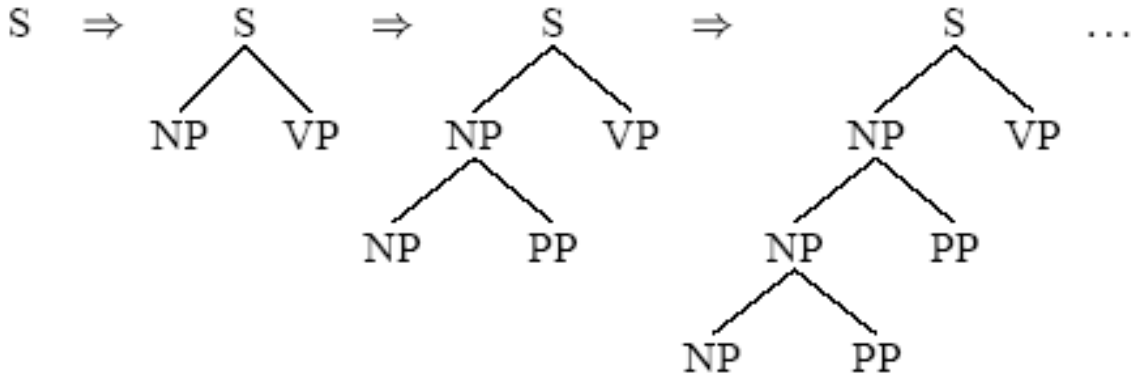




X-vonás mondatszerkezet



Balrekurzió, önbeágyazás



Önbeágyazás balrekurzióval (S → NP VP, NP → Pron S):

0: *A fiú elment.*

1: *A fiú, akit a barátom meghívott, elment.*

2: *A fiú, akit a barátom, akiről a kollégám mesélt, meghívott, elment.*

3: *A fiú, akit a barátom, akiről a kollégám, akivel egy iskolába jártam, mesélt, meghívott, elment.*

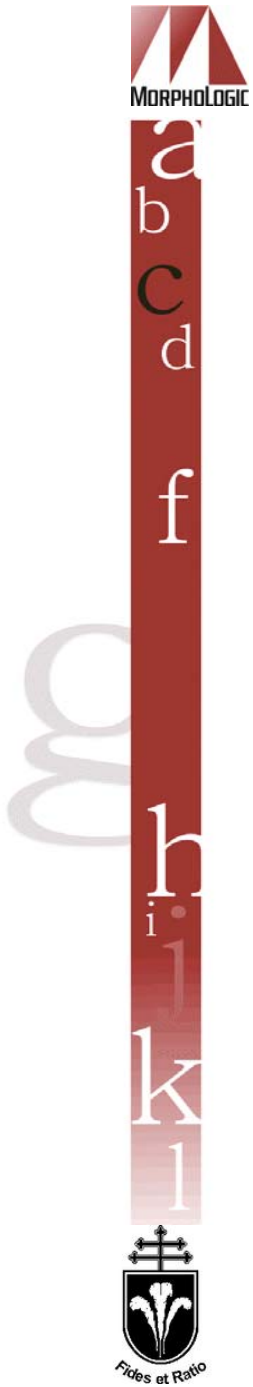
Veremkezelés helyett egyszerű utalás:

Az a fiú elment, akit az a barátom hívott meg, akiről az a kollégám mesélt, akivel egy iskolába jártam.

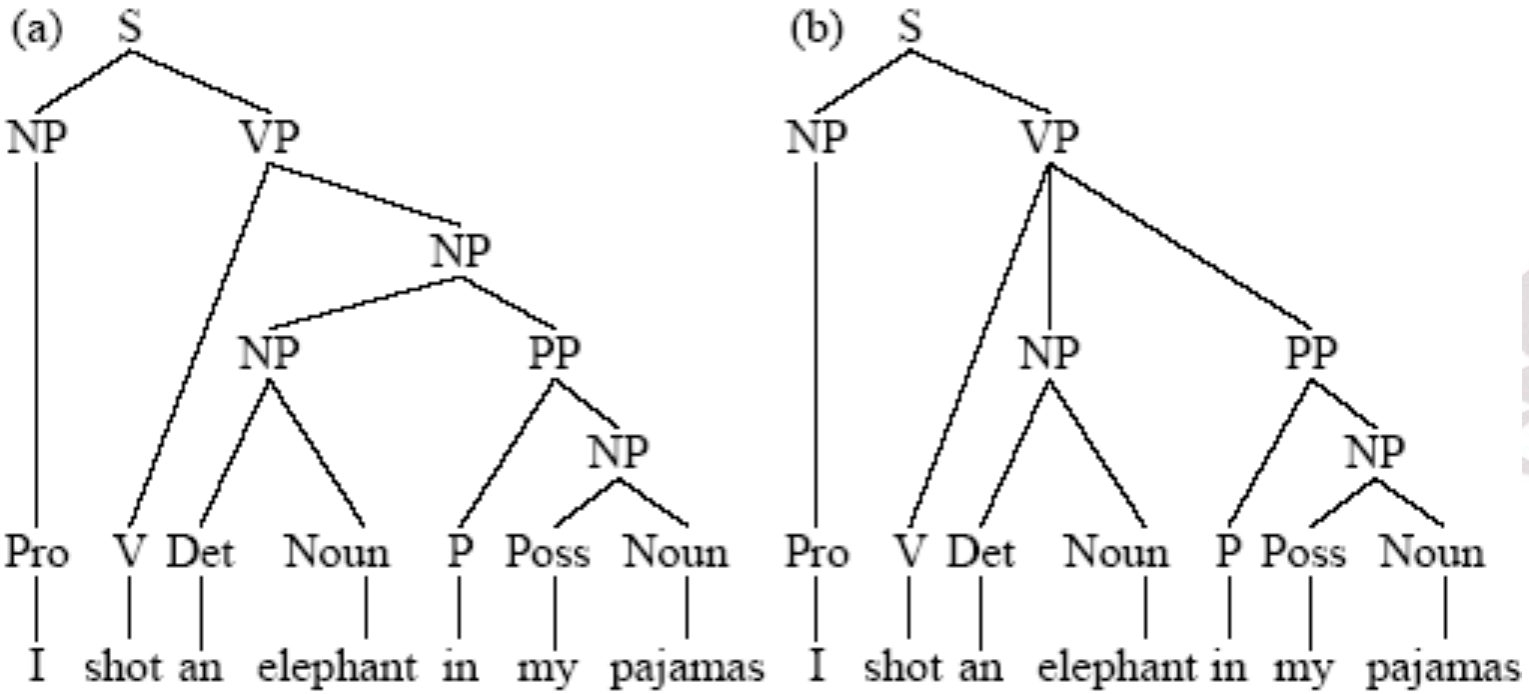
Jobbrekurzió

„az agyag∅
ölelő karjai közül
kibontakozni akaró kocsikerék∅
rettentő nyikorgásától
megriadt juhászkutya∅
bundájába
kapaszkodó kullancs∅
kidülledt félszeméből
alácseppenő könnycseppben
visszatükröződő holdvilág∅
fényétől
illuminált rablólovagvár∅
felvonóhídjából
kiálló vasszegek∅
kohéziós erejének
hatása”

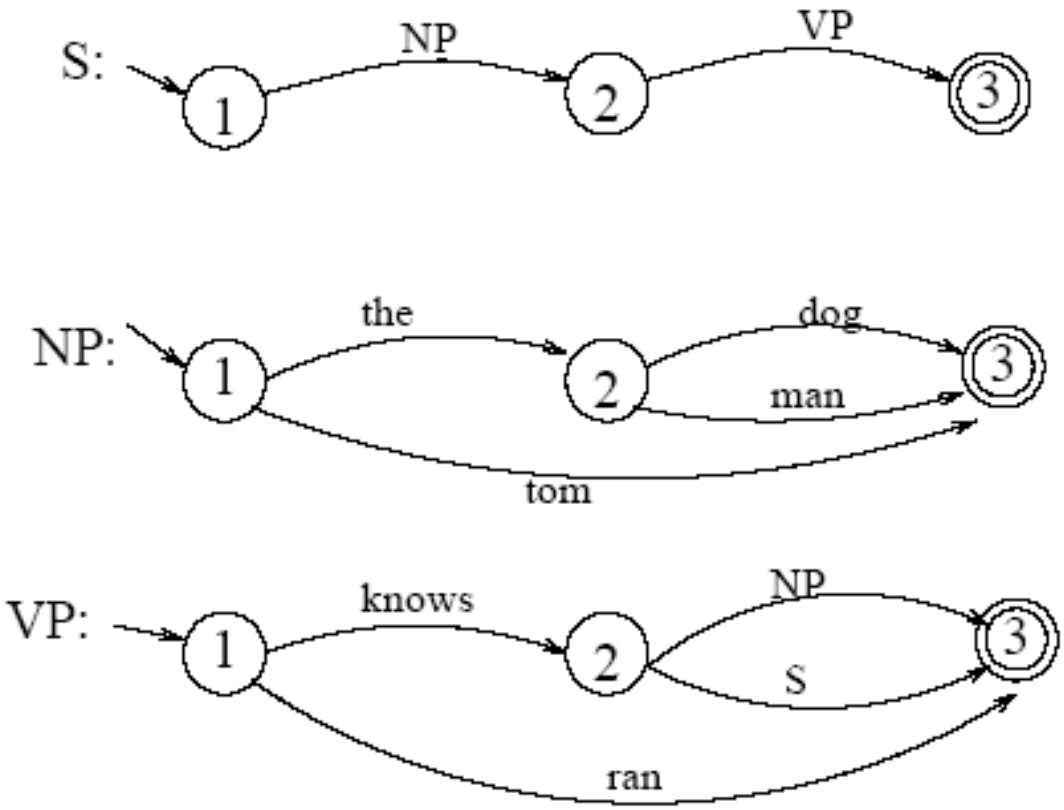
(Fehér G.)



A „PP-attachment” probléma



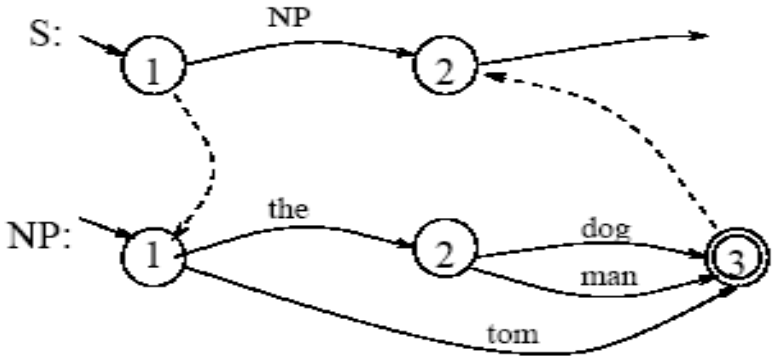
RTN
(Recursive Transition Network)



RTN
(kiegészítések a VÁA-hoz)

A szokásos VÁA működtetésén túl figyelni kell:

- az aktuális bemeneti pozíciót,
- az aktuális állapotot és
- hogy hova kell visszatérni



- összegezve: veremkezelés kell

RTN
(összefoglalva)

- az RTN egymást hívó VÁA-k hálózata: az élek címkéin megjelenik a kategória, azaz más VÁA-k „neve”
- a VÁA (a reguláris nyelvek) *O(n)* idő alatt elemezhetők
- az RTN viszont veremautomata, azaz környezet-független nyelvek elemzésére is alkalmas, tehát csak $O(n^3)$ elemzési idő garantálható

ATN
(az RTN bővítése)

ÉLCÍMKÉK:

WRD *, CAT *, PUSH *, POP, JUMP *

ÉRTÉKEK:

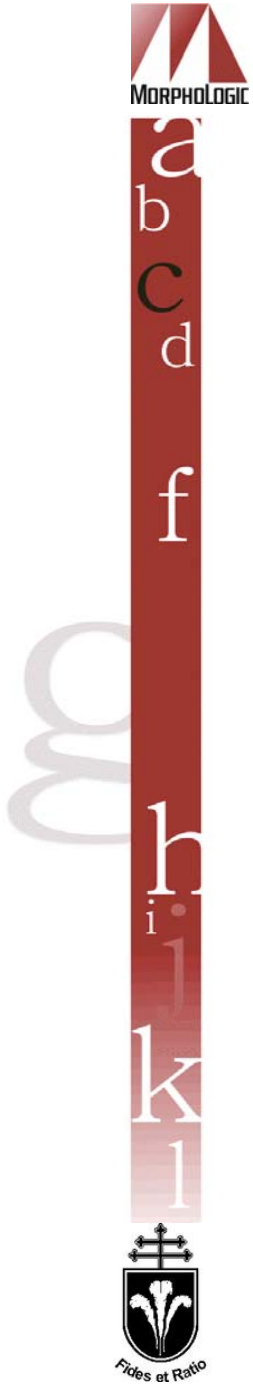
GETR, *, QUOTE, GETF, BUILDQ *, APPEND

TESZTEK:

T, EQ, AND, OR, NOT

AKCIÓK:

SETR, TO



Példák ATN-élekre

(1)
```
<JUMP dest="NP.END">
    <SETR name="tesztregister" value="tesztérték"/>
</JUMP>
```

(2)
```
<CAT name="n" dest="NP.END">
     <SETR name="főnév" value="*"/>
     <SETR name="eset" value="*">
         <GETF name="case"/>
     </SETR>
 </CAT>
```

(3)
```
<WRD name="labdarúgás" dest="NP.END">
    <SETR name="téma" value="sport"/>
</WRD>
```

(4)
```
<PUSH dest="NP">
    <SENDR name="eset"/>
     <JUMP dest="S.VP"/>
</PUSH>
```

(5)
```
<POP>
    <BUILDQ>(det(+)) (n(+)), DET, nomen</BUILDQ>
</POP>
```

(6)
```
<GETF name="személy"/>
```