Nyelvtechnológia 2

BME, 2007. november 13.

Prószéky Gábor

A természetes nyelvek számítógépes ábrázolásának kutatási problémái

- Formális nyelvek a természetes nyelvek kutatásában
- A nyelvmodellek és a nyelv „távolságáról”
- Pontosság és lefedettség
- Túl- és alulgenerálás

Prószéky Gábor

MorphoLogic

Fides et Ratio

A természetes nyelvek modellezésének szintjei és eszközei

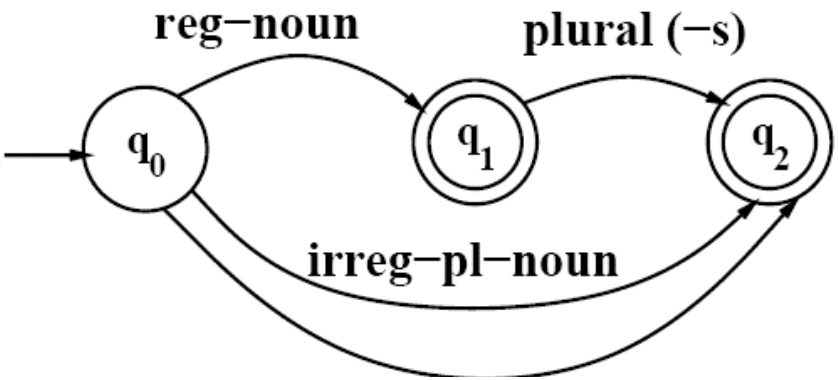
- Nagy paradigmák: a statisztikai, a szabály-alapú és a példa-alapú rendszerek
- A korpusznyelvészet kialakulása: a korpuszok alkalmazása a nyelv különböző szintjeinek kutatásában (treebank)
- A nyelvi kutatások szintjei: fonológia, morfológia, szintaxis, szemantika, pragmatika
- Szövegnyelvészet, dialógus-kutatás, világismeret-kutatás

Az angol morfológia

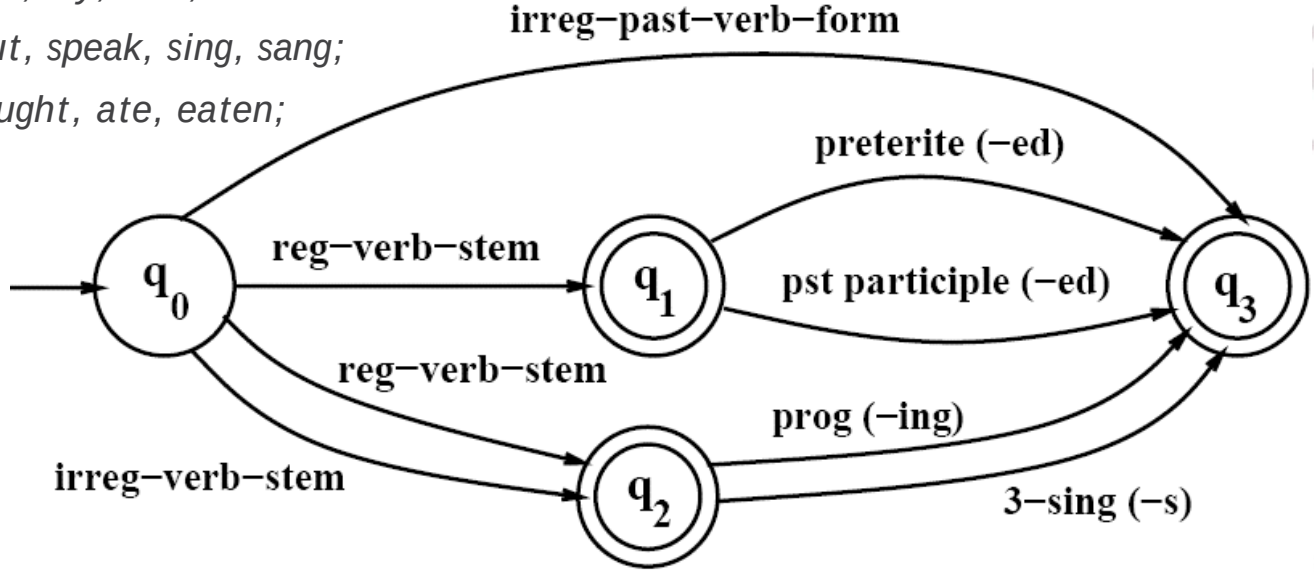
- 1. **walk** (ige): **walk**, **walks**, **walking**, **walked**, **walker**, **walkable**
- 2. **simple** (mn): **simple**, **simpler**, **simplest**, **simply**, **simplier**, **simplicity**
- 3. **computer** (fn; belevéve a képzéssel kapott teljes igei paradigmát is): **computer**, **computer's**, **computers**, **computers'**, **computerize**, **computerizes**, **computerized**, **computerizing**

Az angol morfológia VÁA-ja

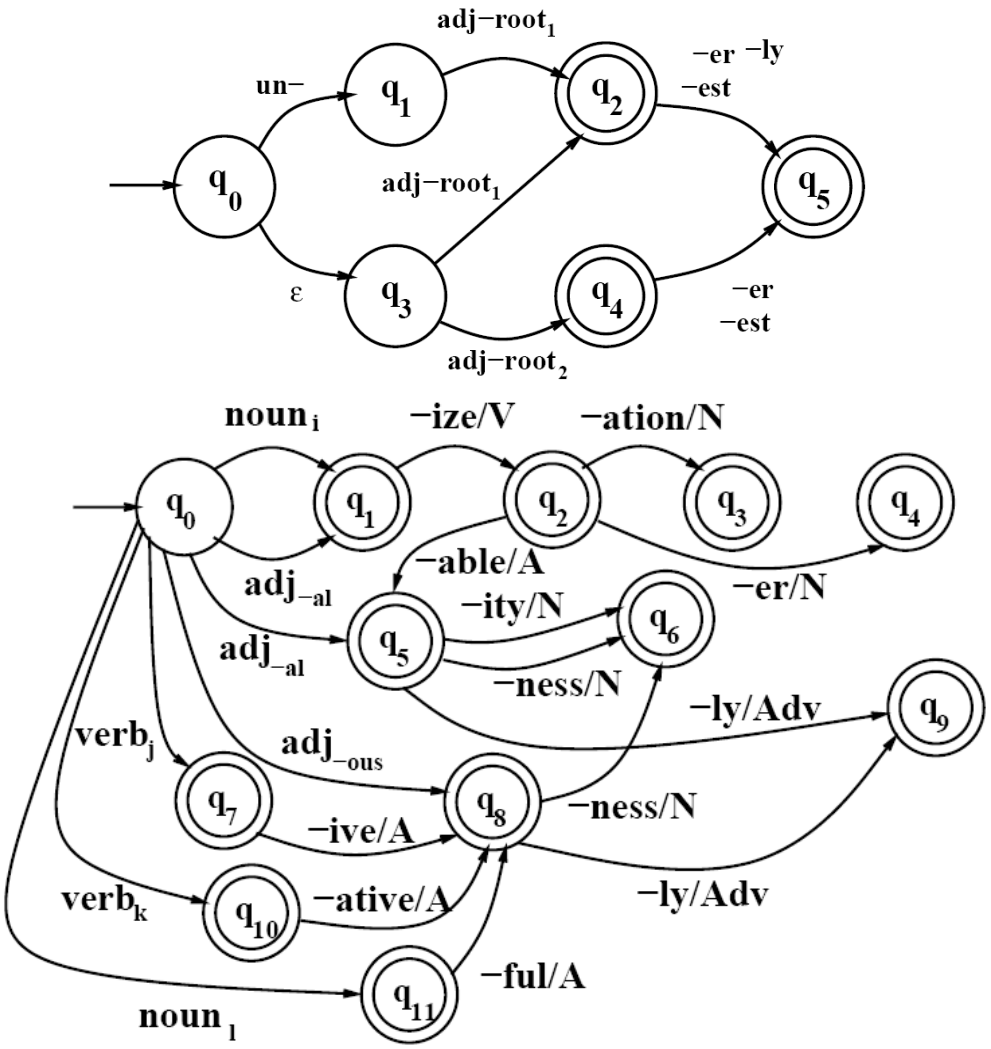
reg-noun: *fox, cat, dog;*
irreg-pl-noun: *geese, sheep, mice;*
irreg-sg-noun: *goose, sheep, mouse*
plural: *-s*



reg-verb-stem: *walk, fry, talk;*
irreg-verb-stem: *cut, speak, sing, sang;*
irreg-past-verb: *caught, ate, eaten;*
past: *-ed;*
past-part: *-ed;*
pres-part: *-ing;*
3sg*: -s*



Az angol morfológia VÁA-ja (2)



A magyar morfológia

- 1. **tesz** (ige): **teszek**, **teszel**, **tesz**, **teszünk**, **tesztek**, **tesznek** stb.... **tettem**, **tetted**, **tette**, **tettük**, **tettétek**, **tették** stb. ... **tenném**, **tennéd**, **tenné**, **tennénk**, **tennétek**, **tennék** stb.... **tehetek**, **tehetsz**, **tehet**, **tehetünk**, **tehettek**, **tehetnek** stb. ... **tevő**, **tevők**, **tevőnek**, **tevőleges** stb. ...
- 2. **egyszerű** (melléknév): **egyszerűen**, **egyszerűt**, **egyszerűnek**, **egyszerűvel**, **egyszerűvé** stb. ... **egyszerűek**, **egyszerűeknek** stb. ... **egyszerűsít**, **egyszerűsödik**, **egyszerűsít** stb. (és az igealakok sora) ... **egyszerűbb**, **egyszerűbbnek**, **egyszerűbbeket** stb.... l**egegyszerűbb**, **legegyszerűbbé** stb.
- 3. **számítógép** (főnév): **számítógépem**, **számítógéped**, **számítógépe** stb. ... s**zámítógépeimet**, **számítógépeidet**, **számítógépeit** stb. ... **számítógépezem**, **számítógépezel**, **számítógépezik** stb. ... **számítógépes**, **számítógépesnek** stb.

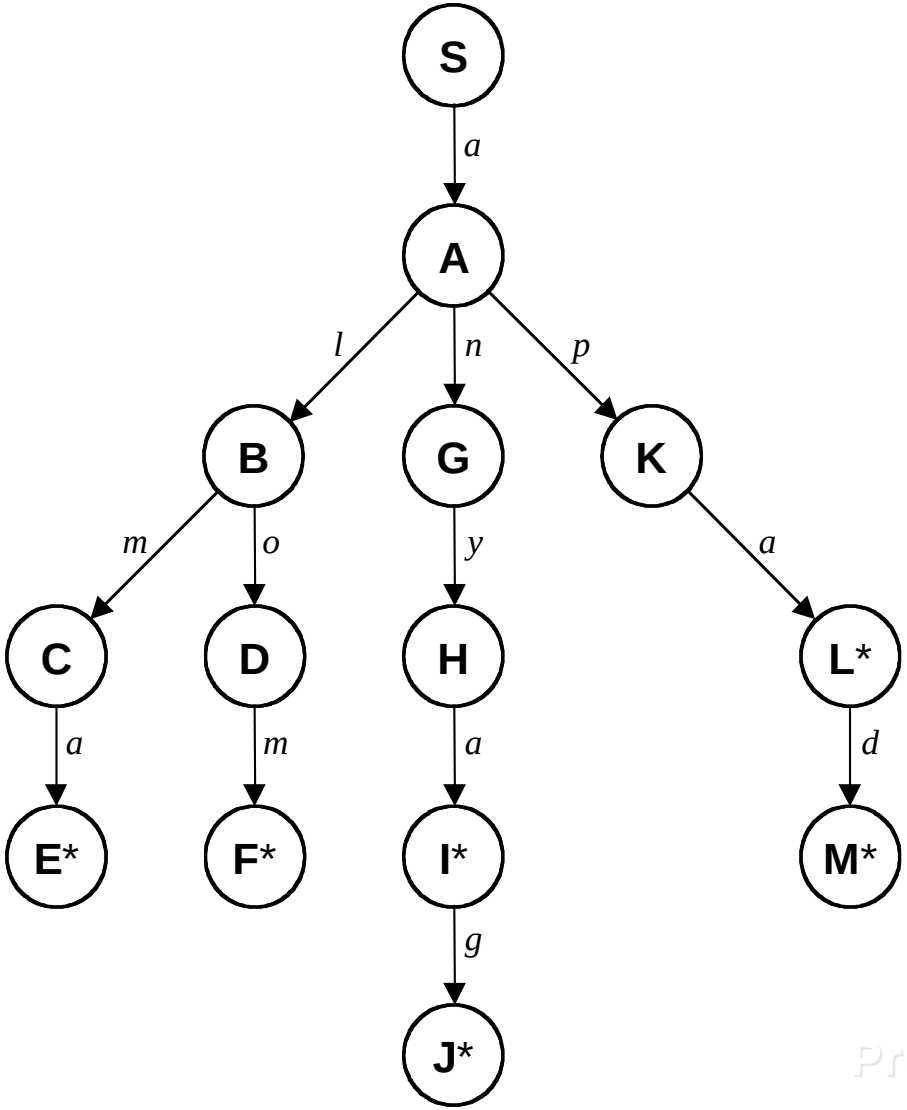
A magyar morfológia - 2

A *morfológiai elemzés* mint program egy olyan *fekete doboz*, mely az alábbi lépéseket végzi el a bemenetül kapott szóalakon:

1. elemi morfémáira bontja;
2. meghatározza a morfémák lexikális alakját;
3. meghatározza az egyes morfémák morfo-szintaktikai tulajdonságait (esetleg más nyelvtani tulajdonságokat is)

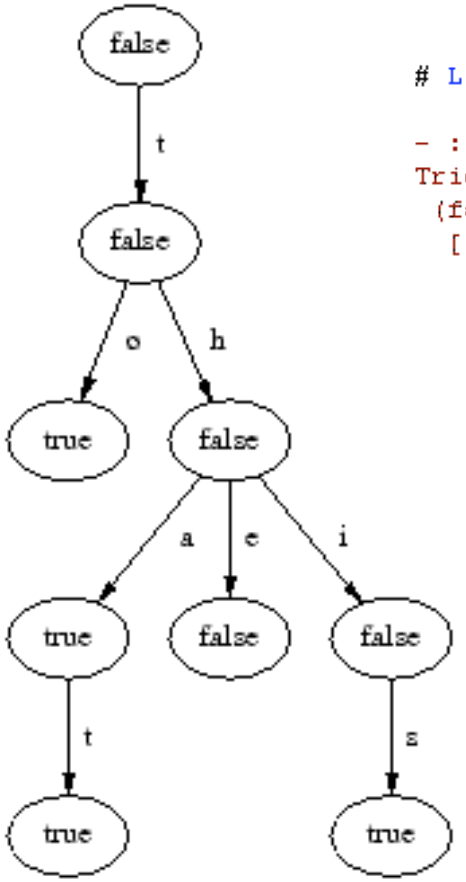
Átmenetgráfos ábrázolás

(alma, alom, anya, anyag, apa, apad)



Trie (=szófa)
(to, this, the, that)

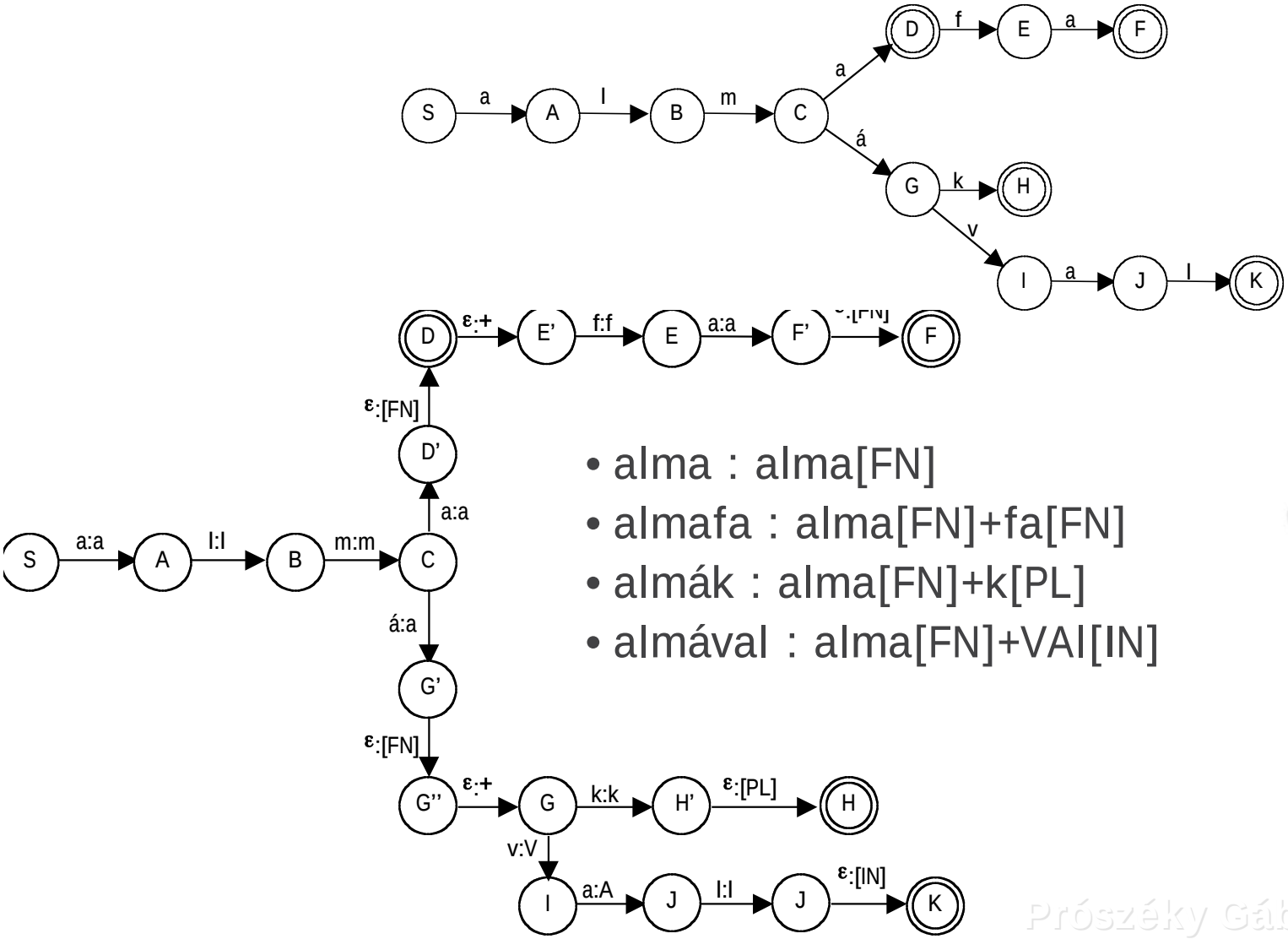
A szófa egy olyan, a szavak rákövetkező karaktereivel címkézett élsorozatokat tartalmazó fa, amelyben egy szót úgy találunk meg, hogy végigjárjuk karakterenként.




```
# List.fold_left (fun x y -> enter (explode y) x)
                         emptytrie ["to";"this";"the";"that"];;
- : trie =
Trie
 (false,
  [('t',
    Trie
     (false,
      [('h',
        Trie
         (false,
          [('a', Trie (false, [('t', Trie (true, []))]));
           ('e', Trie (true, []));
           ('i', Trie (false, [('s', Trie (true, []))]))]));
       ('o', Trie (true, []))]))])
```

Szófa és véges fordító (transducer)

(alma, almafa, almák, almával)



- alma : alma[FN]
- almafa : alma[FN]+fa[FN]
- almák : alma[FN]+k[PL]
- almával : alma[FN]+VAl[IN]

A módosított szófa

(alma, alom, anya, anyag, apa, apad, aránytalanság)

- Ha tudjuk, hogy véges sok elemünk van, módosítható az elágazási helyeknél:
  alm–a
  alo–m
  anya
  anyag
  apa
  apad
  ar–ánytalanság
- Akkor éri meg, ha jelentősen különböznek a szóvégek
- További módosítások: az előtagok (igekötők, *re-*, *pre-*, *anti-* stb.) elkülönítése mellett a tipikus és ritka kezdő betűpárok egyedi kódolása
- Az angol lexikonok tanúsága szerint 262=676 indító betűpárból csak 309 létezik, amiből 88 csak 15-nél kevesebb szó elején)

A Kay-féle szótárábrázolás
(alma, alom, anya, anyag, apa, apad, aránytalanság)

- Kay (1977): tömörítés numerikus prefixekkel

  alma – 0

  alom – 2

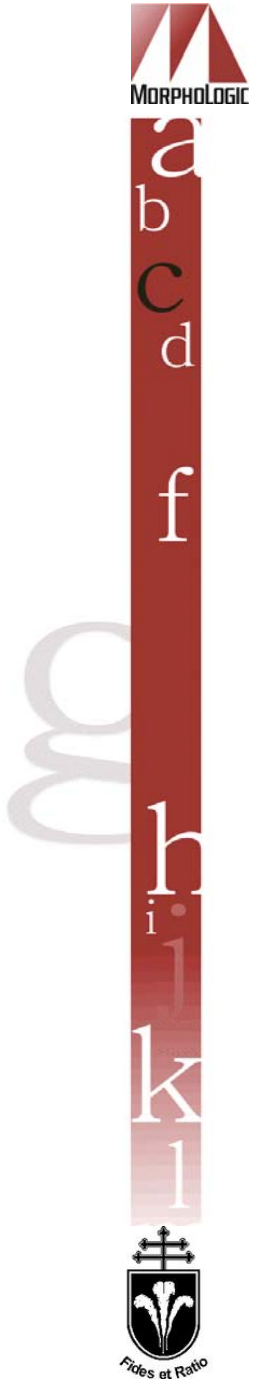
  anya – 1

  anyag – 4

  apa – 1

  apad – 3

  aránytalanság – 1
- Tehát a szótár:
  alma, 2om, 1nya, 4g, 1pa, 3d, 1ránytalanság
- Akkor éri meg, ha hasonlítanak a szókezdetek
  (nagy szótár esetén mindig!)



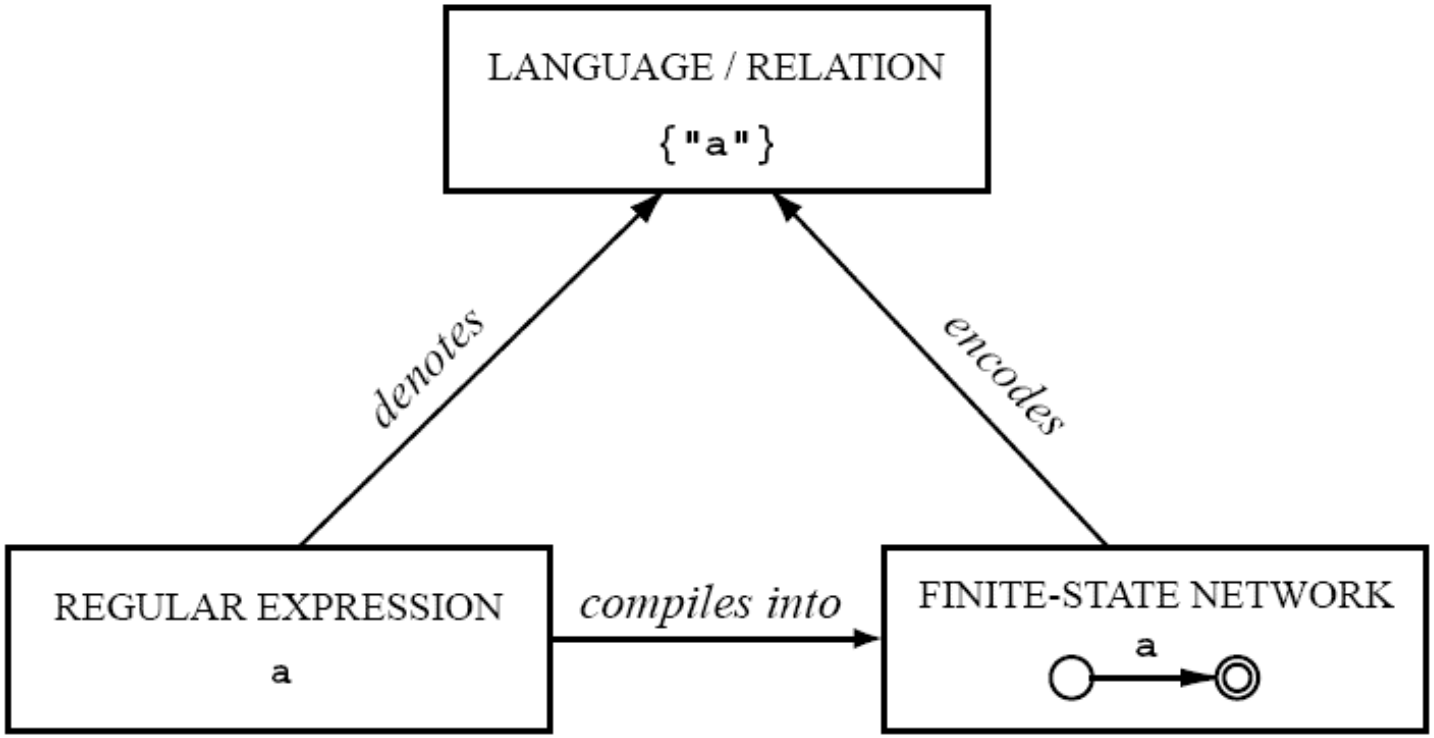
A morfológiai elemzéshez kapcsolódó alapfogalmak

- szókészlet
- szótárábrázolás
- keresési lépések
- túlgenerálás
- zártság

Emlékezetfrissítés

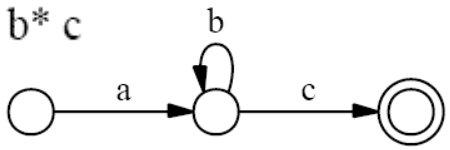
- ✓ **Nyelv**:
  füzérek halmaza
- ✓ **Reguláris nyelv**:
  füzérek olyan halmaza, mely konkatenációval, iterációval és egyszerű halmazműveletekkel hozható létre
- ✓ **Reguláris kifejezés**:
  a reguláris nyelvet leíró kompakt formula
- ✓ **Véges állapotú automata**:
  egy olyan absztrakt gép, mely egy reguláris nyelvet fogad el

Reguláris kifejezés – nyelv – VÁA

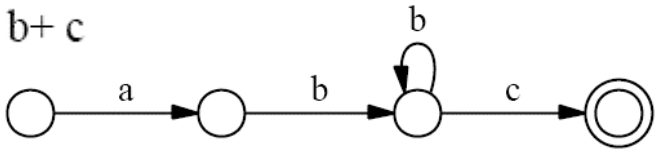


Reguláris kifejezések VÁA-ként

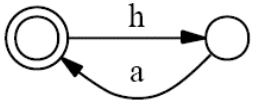
a b* c



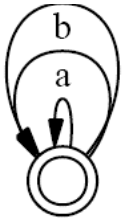
a b+ c



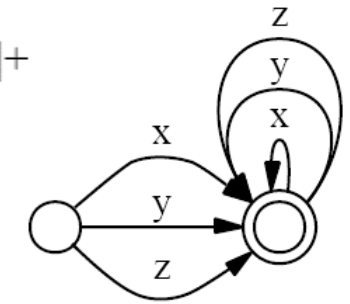
[h a]*



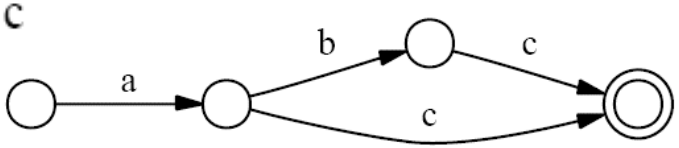
[a | b | c]*



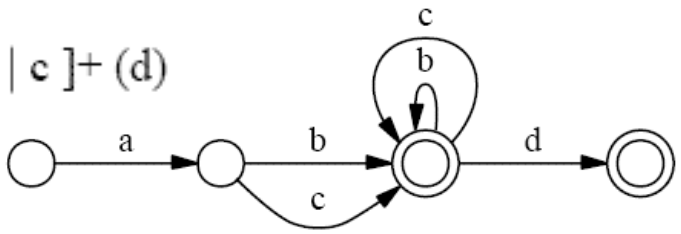
[x | y | z]+



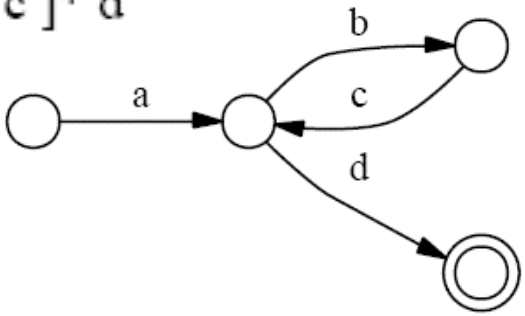
a (b) c



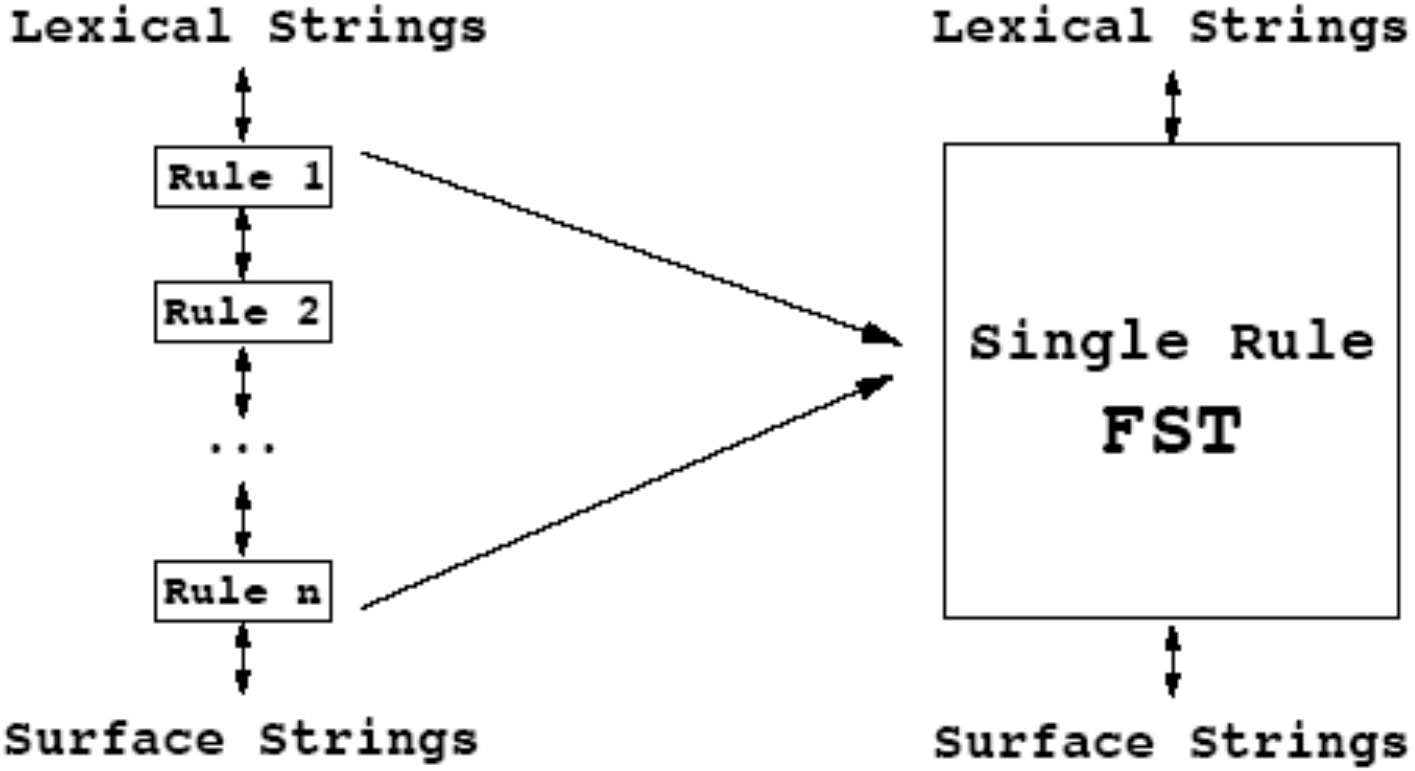
a [b | c]+ (d)



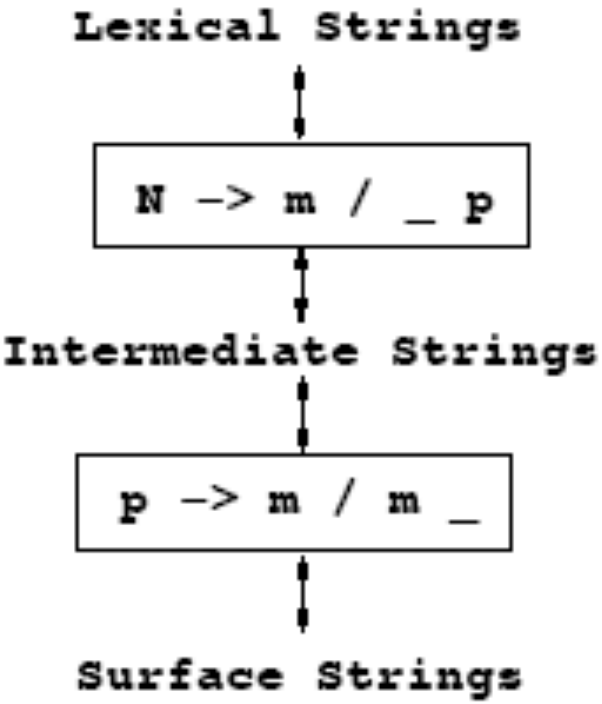
a [b c]* d

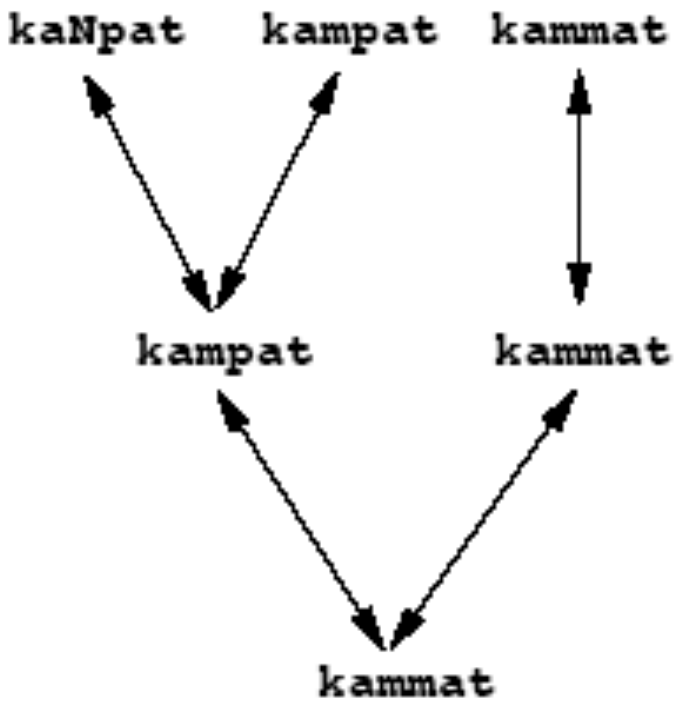


Újraírószabályok egy VÁA-ban

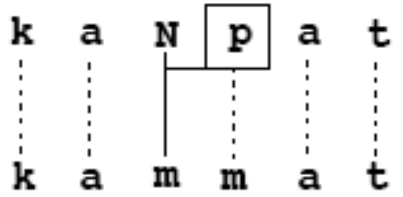


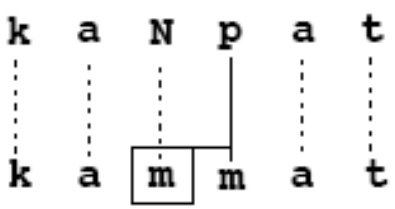
Újraírószabályok egy VÁA-ban - 2

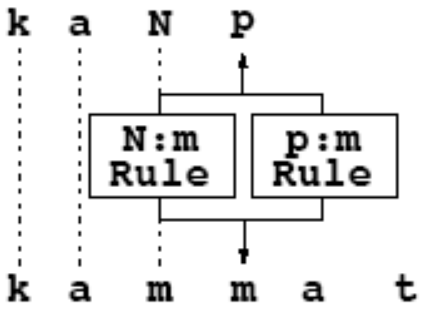


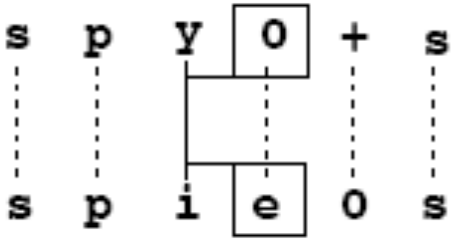


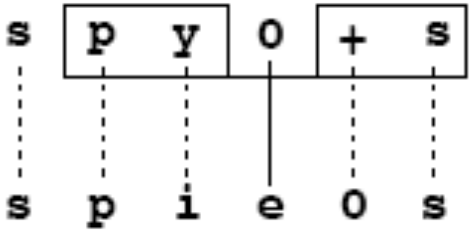
Kétszintes megfogalmazások



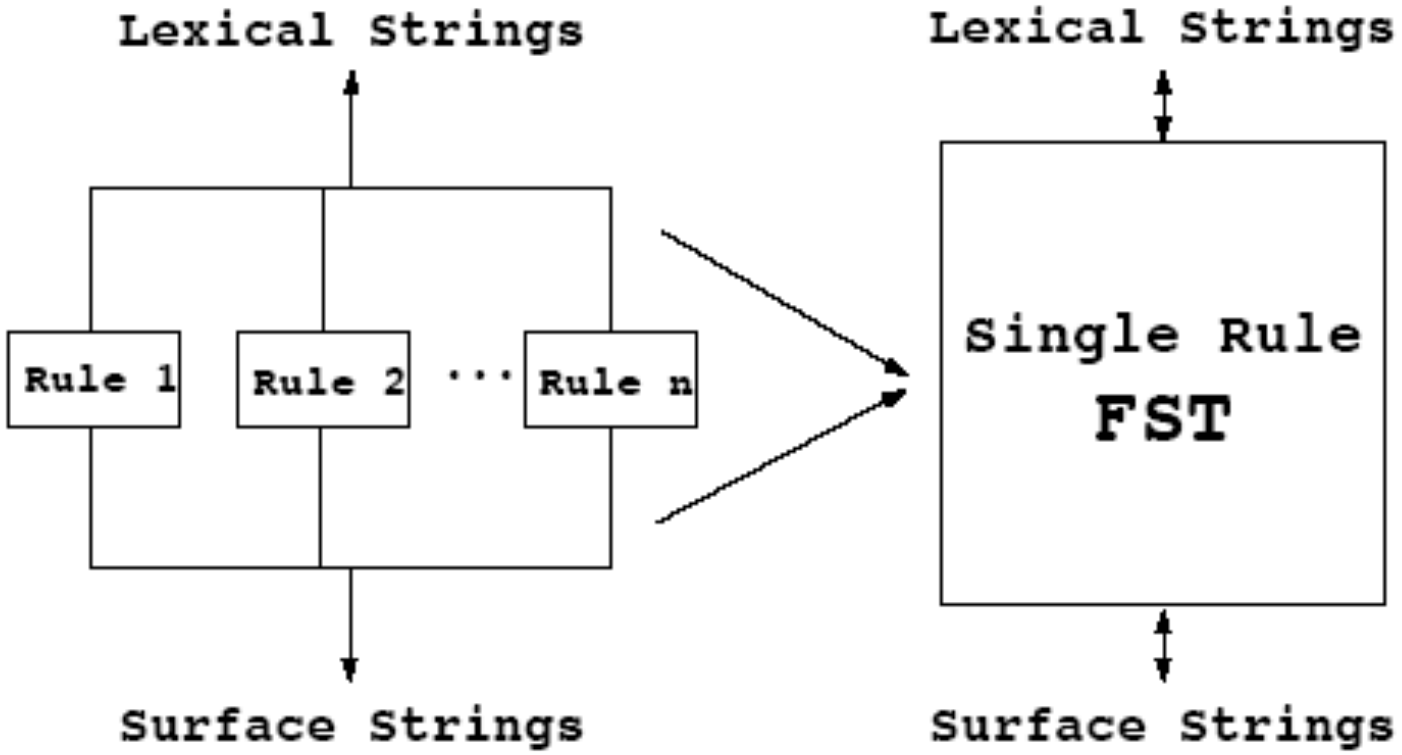








Párhuzamosság: VÁA-metszet



Két szint: felszíni és lexikális

m	o	v	e	+	e	d	*lexical*
m	o	v	0	0	e	d	*surface*

$$X : X \Rightarrow \_\_ \quad (1)$$

$$+ : 0 \Rightarrow \_\_ \quad (2)$$

$$e : 0 \Rightarrow v : v \_\_\_ + : 0 \quad (3)$$

A kétszintes szabályok

L:S => E

„Csak akkor, de nem mindig."

L csak az E környezetben realizálódik S-ként.

Az S-ként realizált L nem megengedett a ¬E környezetben.

Ha L:S, akkor annak E környezetben kell lennie.

Persze L:¬S is engedélyezett lehet az E környezetben.

L:S <= E

„Mindig, de nem csak akkor."

L mindig S-ként realizálódik az E környezetben.

Az ¬S-ként realizált L nincs megengedve az E környezetben.

Ha L illeszkedik az E környezetbe, akkor L:S.

Persze L:S előfordulhat máshol is.

A kétszintes szabályok (2)

L:S <=> E

„Akkor és csak akkor"

Az L S-ként akkor és csak akkor realizálódik, ha E a környezet.

Mind L:S => E, mind L:S <= E fennáll.

L:S kötelező az E környezetben.

L:S sehol máshol nem fordulhat elő.

L:S /<= E

„Soha."

L soha nem realizálódik S-ként az E környezetben.

Az S-ként realizált L nincs megengedve az E környezetben.

Ha L az E környezetben áll, akkor fenn kell álljon L:¬S.

Egy konkrét kétszintes szabály

```
t:c => ___ i
```

```
t:c <= ___ i
```

```
t:c <=> ___ i
```

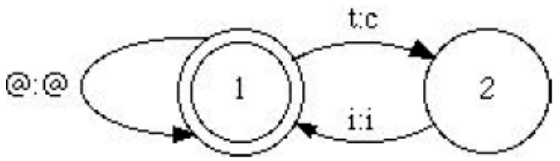
```
t:c /<= ___ i:ê
```

```
-
  t i @
  c i @
  -----
1: 2 1 1
2. 0 1 0
```

```
  t t i @
  c t i @
  -------
1:  1 2 1 1
2:  1 2 0 1
```

```
-
  t t i @
  c @ i @
  -------
1:  3 2 1 1
2:  3 2 0 1
3.  0 0 1 0
```

```
  t i @
  c ê @
  -----
1:  2 1 1
2:  2 0 1
```



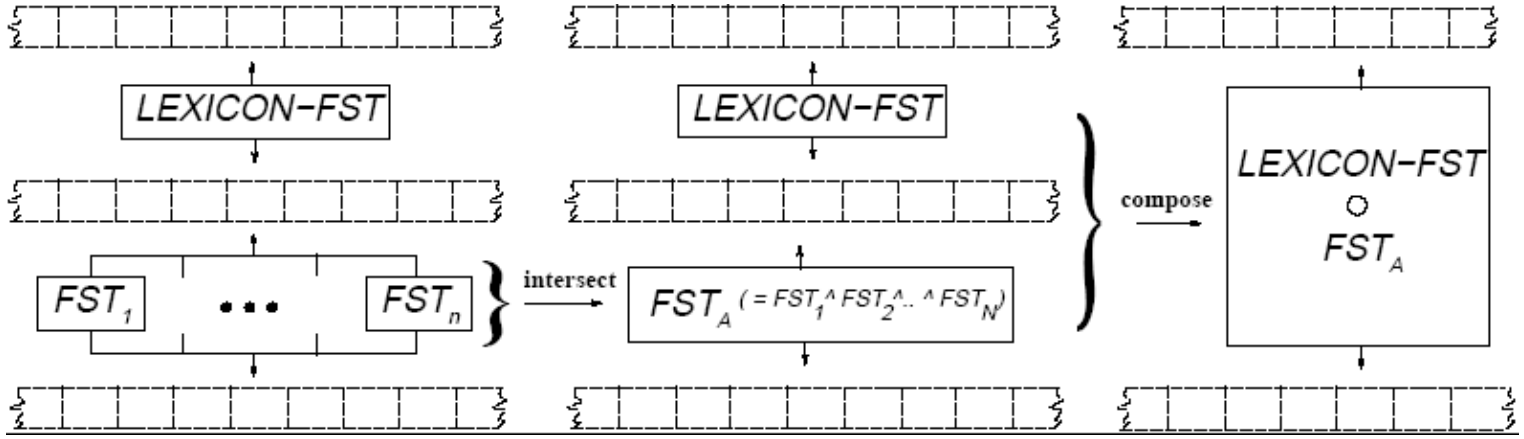
A kétszintes rendszer

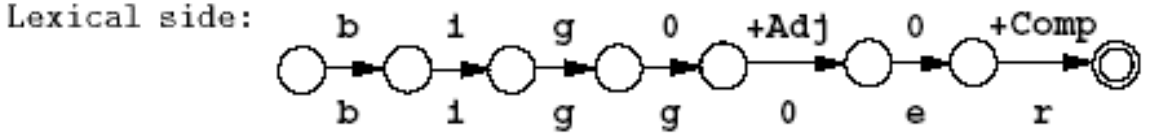
- a felhasználó környezetfüggő szabályokat ír
- minden jelenségre egy szabály (a többi a rendszer dolga)
- az ábécé(k) megadandó(k):

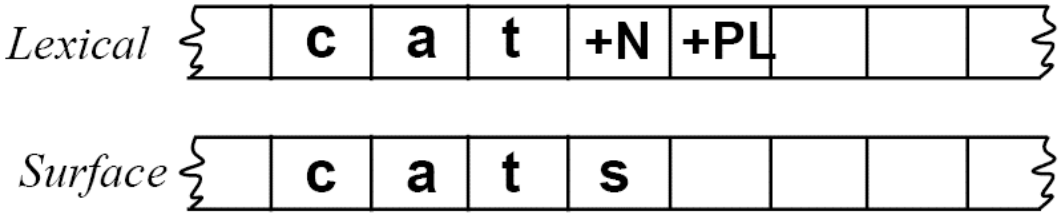
```
SUBSET C     p t k b d g m n ng s l r w y
SUBSET V     i e a o u
SUBSET S     p t k b d g
SUBSET NAS   m n ng
```

- lexikonok és folytatási osztályok
- metakarakterek használhatók
- speciális szimbólumok (üres, akármi)
- szabályfordító és táblázatos forma

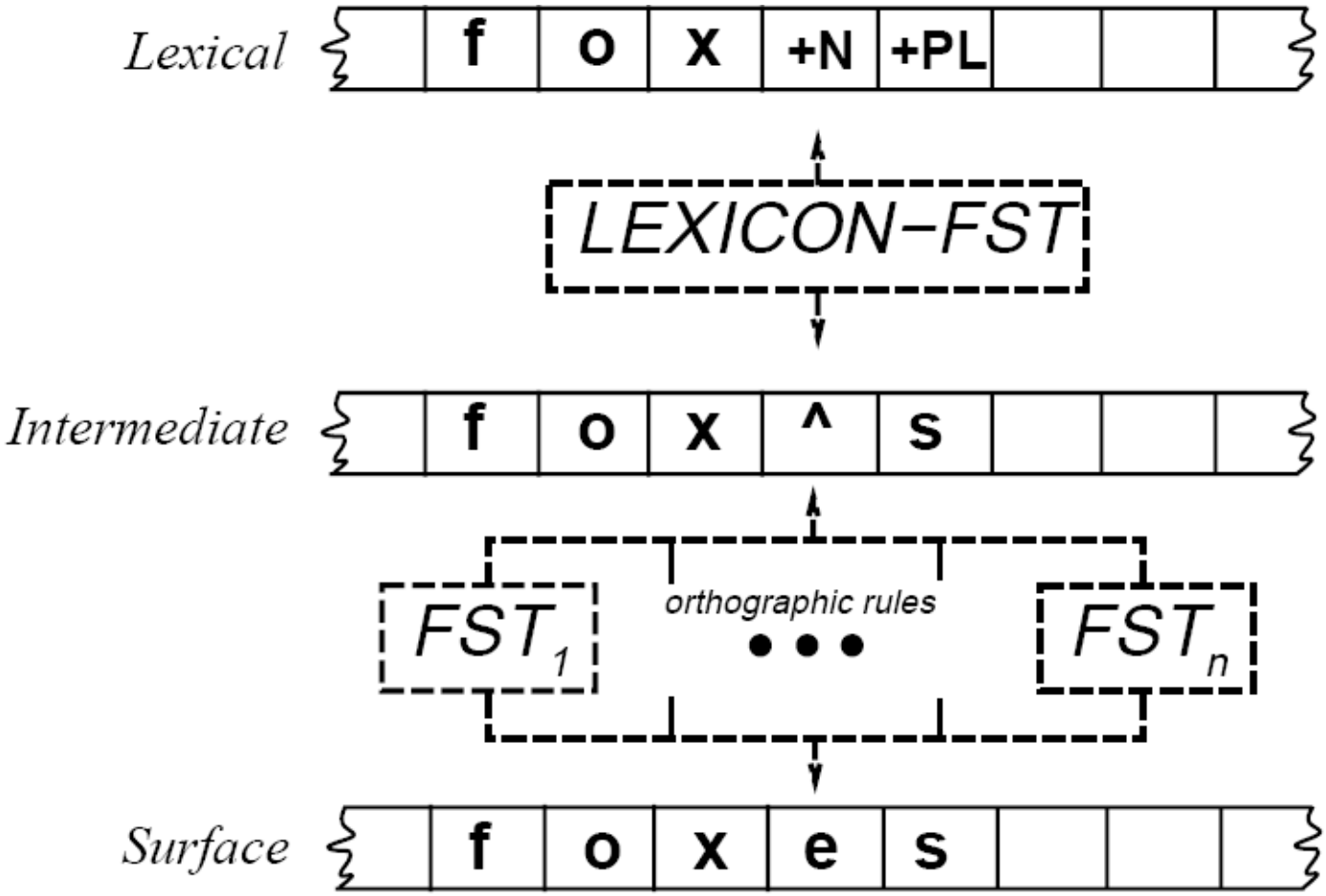
Később: szabályok és lexikonok kompozíciója



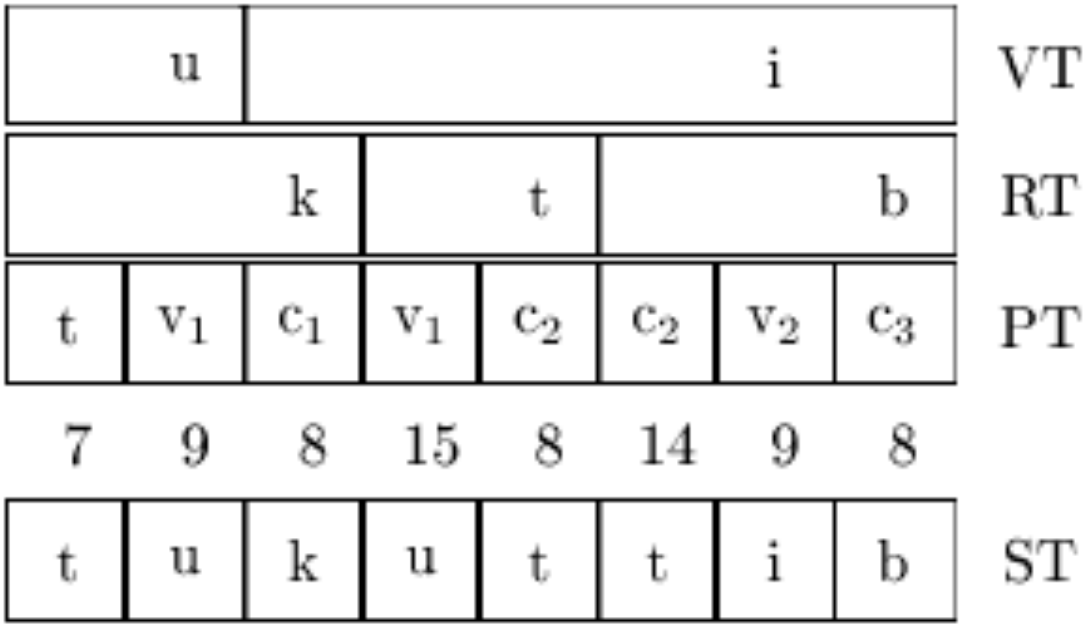




Szabályok és lexikonok metszete a gyakorlatban

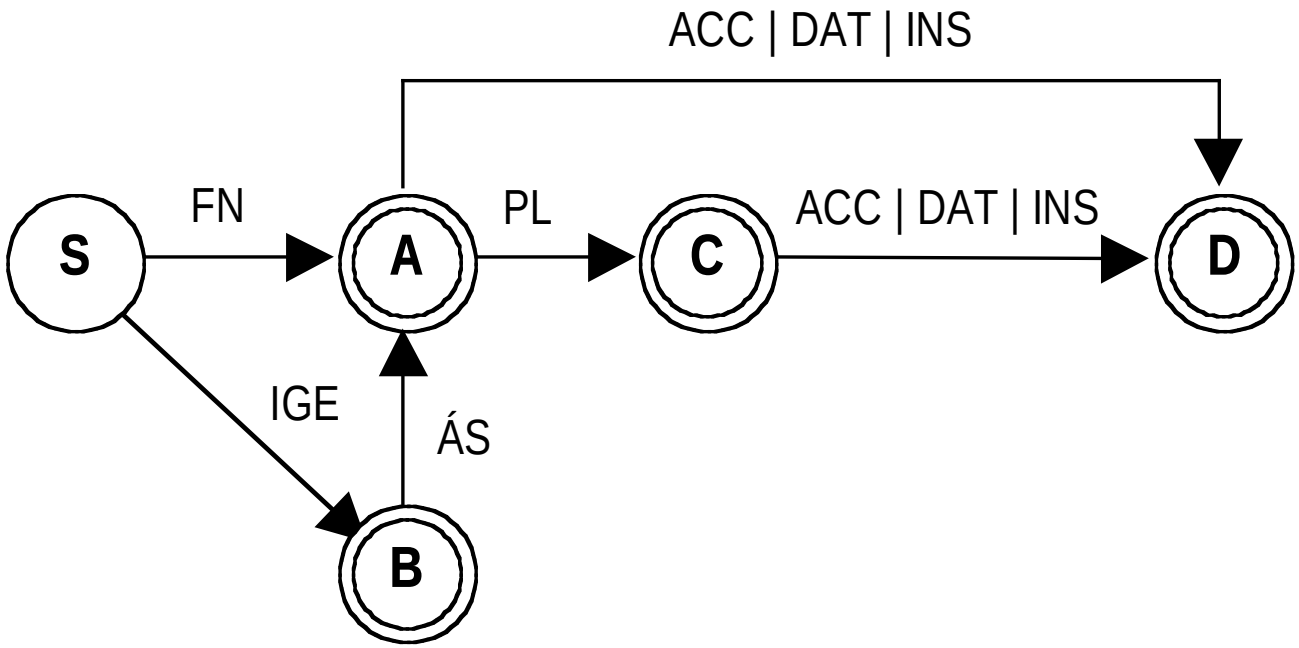


Több szalag: felszíni és több lexikális

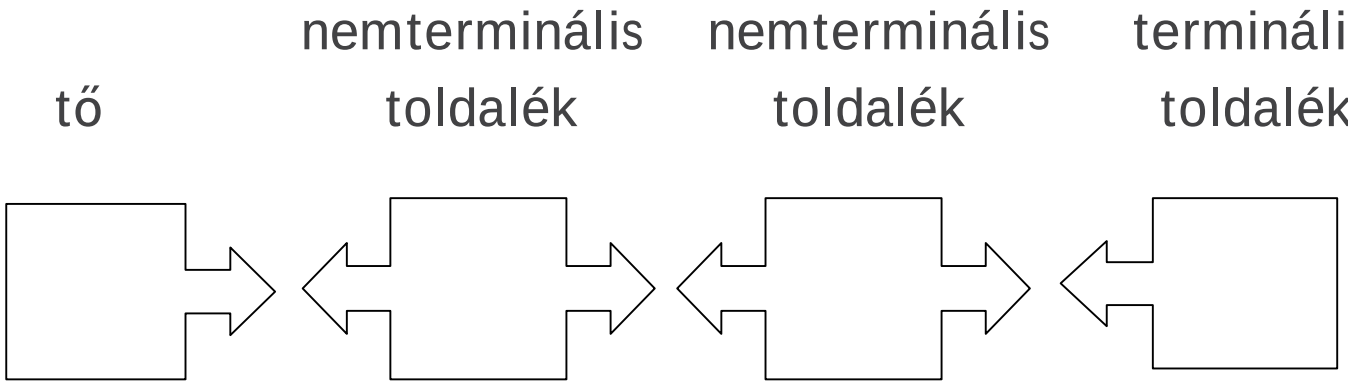


Folytatási osztályok

Leegyszerűsített magyar névszói toldalékolás:



Szóalaktani alapséma



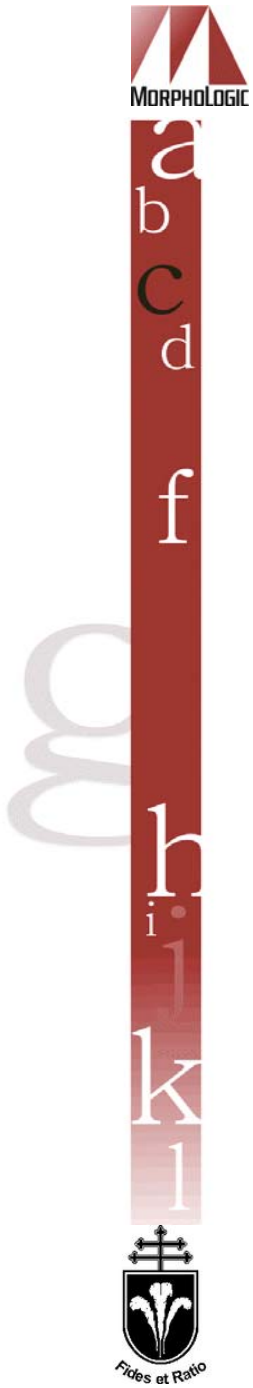
(relatív) tő / relatív toldalék

relatív tő / relatív toldalék

relatív tő / (relatív) toldalék

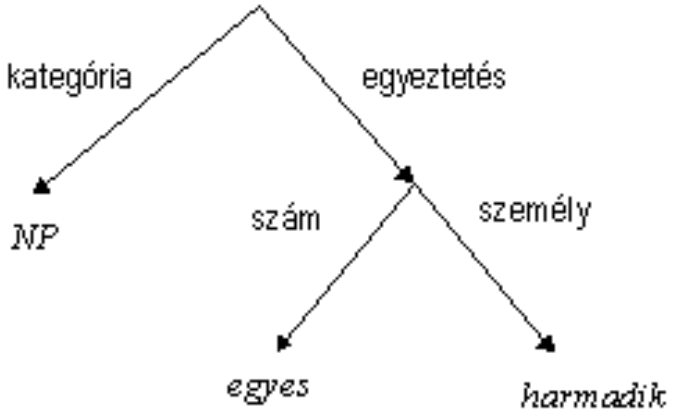
HUMOR

- High-speed Unification Morphology
- folytatási osztályok (mátrix)
- jegy-érték párok
- unifikáció: részletes definíció később
- ld. Prolog, de nem rögzített aritás
- unifikáció vs. unifikálhatóság
- minden tulajdonság jegyként
- nincs más „valós” művelet, csak az unifikálhatóság-ellenőrzés



Jegyszerkezetek

$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} szám = egyes \\ személy = harmadik \end{bmatrix} \end{bmatrix}$$



Unifikáció

$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} szám = egyes \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} személy = harmadik \end{bmatrix} \end{bmatrix}$$

A két jegyszerkezet unifikációjának eredménye:

$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} szám = egyes \\ személy = harmadik \end{bmatrix} \end{bmatrix}$$

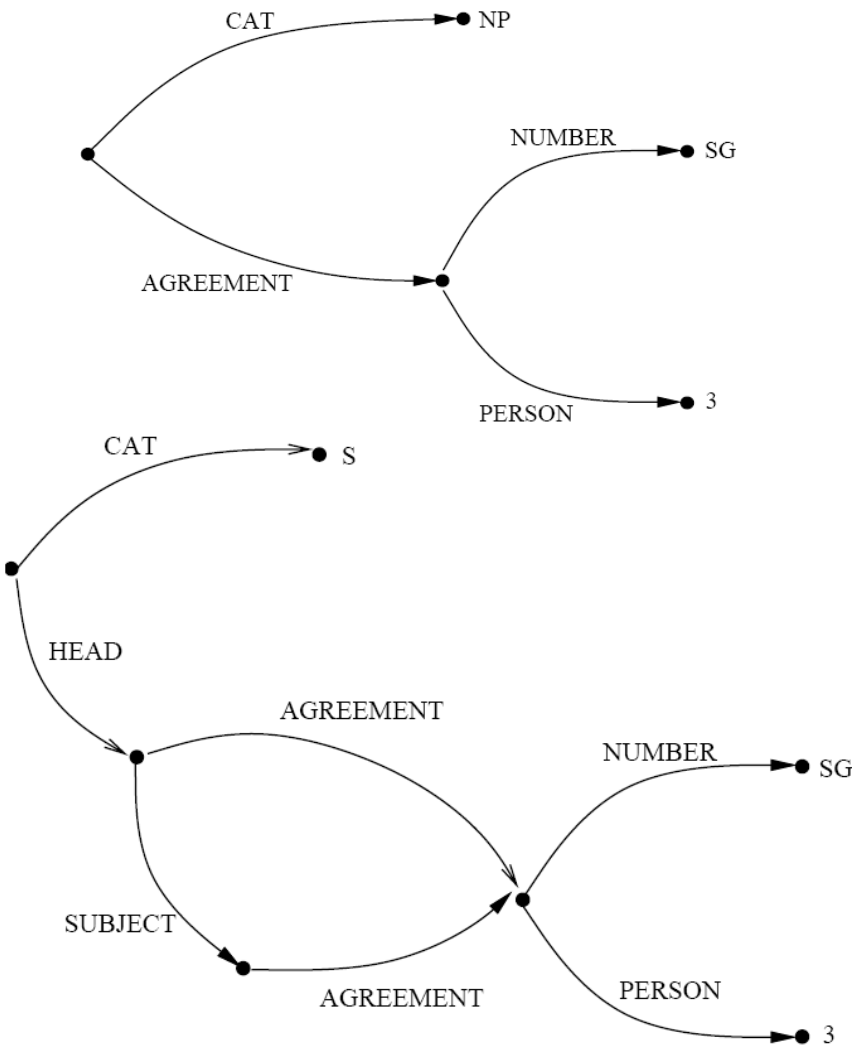
$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} szám = egyes \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} kategória = & NP \\ egyeztetés = & \begin{bmatrix} szám = többes \end{bmatrix} \end{bmatrix}$$

MorphoLogic

Fides et Ratio

DAG-ok ábrázolása



Az unifikáció definíciója

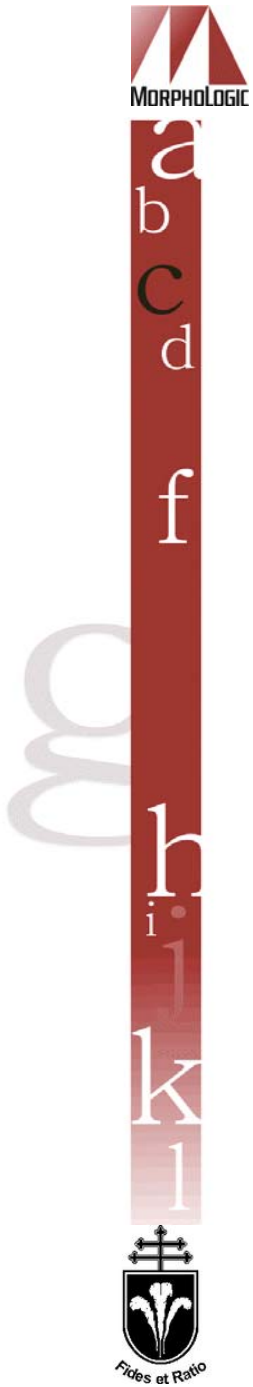
```
function UNIFY(f1, f2) returns fstructure or failure

 f1-real ← Real contents of f1
 f2-real ← Real contents of f2

 if f1-real is null then
   f1.pointer ← f2
   return f2
 else if f2-real is null then
   f2.pointer ← f1
   return f1
 else if f1-real and f2-real are identical then
   f1.pointer ← f2
   return f2
 else if both f1-real and f2-real are complex feature structures then
   f2.pointer ← f1
   for each feature in f2-real do
     other-feature ← Find or create
                          a feature corresponding to feature in f1-real
     if UNIFY(feature.value, other-feature.value) returns failure then
       return failure
   return f1
 else return failure
```

Bináris kérdések
a magyar morfo-fonológiáról

		α = +	α = –
1	α névszó	névszó	ige
2	α fn	főnév	melléknév, számnév
3	α szótári	szótári alapalak	nem szótári alapalak
4	α elöl	elöl képzett	hátul képzett
5	α kerek	ajakkerekítéses	nem ajakkerekítéses
6	α PL	többes szám	nem állhat többes számban
7	α PLkötő	PL kötőhanggal	PL nem kötőhanggal
8	α PERS	birt. szem.ragos	nem kap birt. szem.ragot
9	α ACC	van tárgyesete	nem tárgyesetes
10	α ACCkötő	ACC kötőhanggal	ACC nem kötőhanggal
11	α DAT	van részesesete	nincs részesesete
12	α INS:ß	van eszk.h.esete	nincs eszk.h.esete
13	α ÁS	*-ás/-és* képzős	nem kap *-ás/-és* képzőt



Szótövek tára

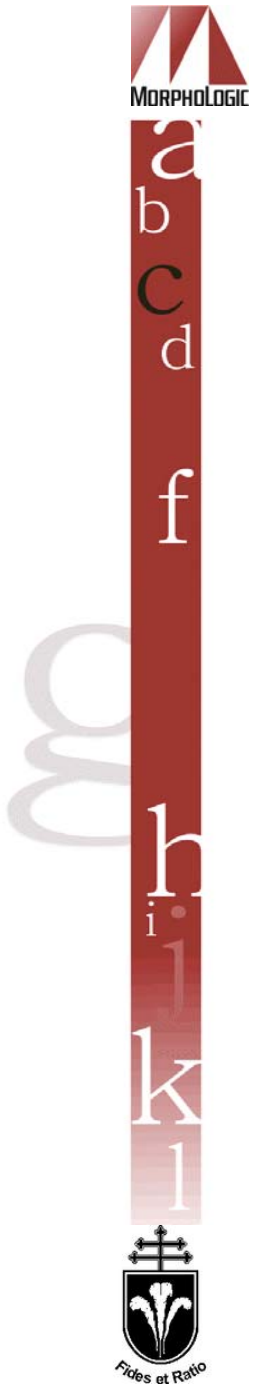
szó []
[+névszó +fn +szótári –elöl –kerek –PL –PERS +ACC –ACCkötő +DAT +INS:V]

szav []
[+névszó +fn –szótári –elöl –kerek +PL +PLkötő +PERS –ACC +DAT –INS]

képez []
[–névszó +szótári +elöl –kerek –ÁS]

képz []
[–névszó –szótári +elöl –kerek +ÁS]

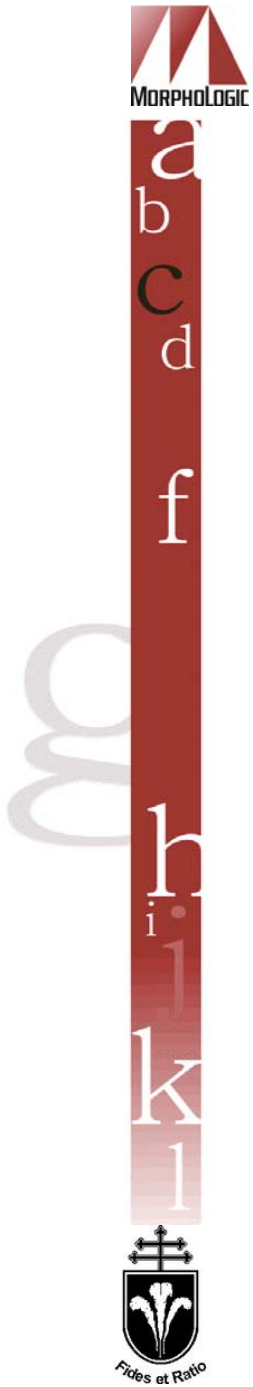
…



Toldalékok tára

ás	[–névszó –elöl +ÁS]
	[+névszó +fn +szótári –elöl –kerek +PL +PLkötő +ACC –ACCkötő +DAT +INS:S]
és	[–névszó +elöl +ÁS]
	[+névszó +fn +szótári +elöl –kerek +PL +PLkötő +ACC –ACCkötő +DAT +INS:S]
ak	[+névszó –elöl –kerek +PL +PLkötő]
	[+névszó –elöl –kerek –PL –PERS +ACC +ACCkötő +DAT +INS:K]
ek	[+névszó +elöl –kerek +PL +PLkötő]
	[+névszó +elöl –kerek –PL –PERS +ACC +ACCkötő +DAT +INS:K]
nak	[+névszó –elöl +DAT]
	[]
nek	[+névszó +elöl +DAT]
	[]

...



Unifikációs morfológia

szó	[+névszó +fn **+szótári** –elöl –kerek –PL –PERS +ACC –ACCkötő +DAT +INS:V]
**szav*	[+névszó +fn **-szótári** –elöl –kerek +PL +PLkötő +PERS –ACC +DAT –INS]
szó+nak	[**+névszó** +fn +szótári **-elöl** –kerek –PL –PERS +ACC –ACCkötő **+DAT** +INS:V] [**+névszó -elöl +DAT**]
**szav+nak*	[**+névszó** +fn –szótári **-elöl** –kerek +PL +PLkötő +PERS –ACC **-DAT**] [**+névszó -elöl +DAT**]
**szó+vel*	[**+névszó** +fn +szótári **-elöl** –kerek –PL –PERS +ACC –ACCkötő +DAT **+INS:V**] [**+névszó +elöl +INS:V**]
**szav*	[+névszó +fn **-szótári** –elöl –kerek +PL +PLkötő +PERS –ACC +DAT –INS]
képz+és+nek	[**-névszó** –szótári **+elöl** –kerek **+ÁS**] [**-névszó +elöl +ÁS**] [**+névszó** +fn +szótári **+elöl** –kerek +PL +PLkötő +ACC –ACCkötő **+DAT** +INS:S] [**+névszó +elöl +DAT**]

MorphoLogic

Fides et Ratio