Egy szónak is száz a vége

Oravecz Csaba

MTA Nyelvtudományi Intézet
Korpusznyelvészeti Osztály
oravecz@nytud.hu

Magyar tudomány napja, MTA, 2003. 11. 04.

Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

  karaktersorozatokat

- |a|z|␣|a|v|a|r|␣|s|í|r|: 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező jelek

Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

  karaktersorozatokat

- |a|z|␣|a|v|a|r|␣|s|í|r|: 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező jelek

  Az őszi avar sír a lába alatt.

Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

  karaktersorozatokat

- |a|z|␣|a|v|a|r|␣|s|í|r|: 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező jelek

  Az őszi avar sír a lába alatt.

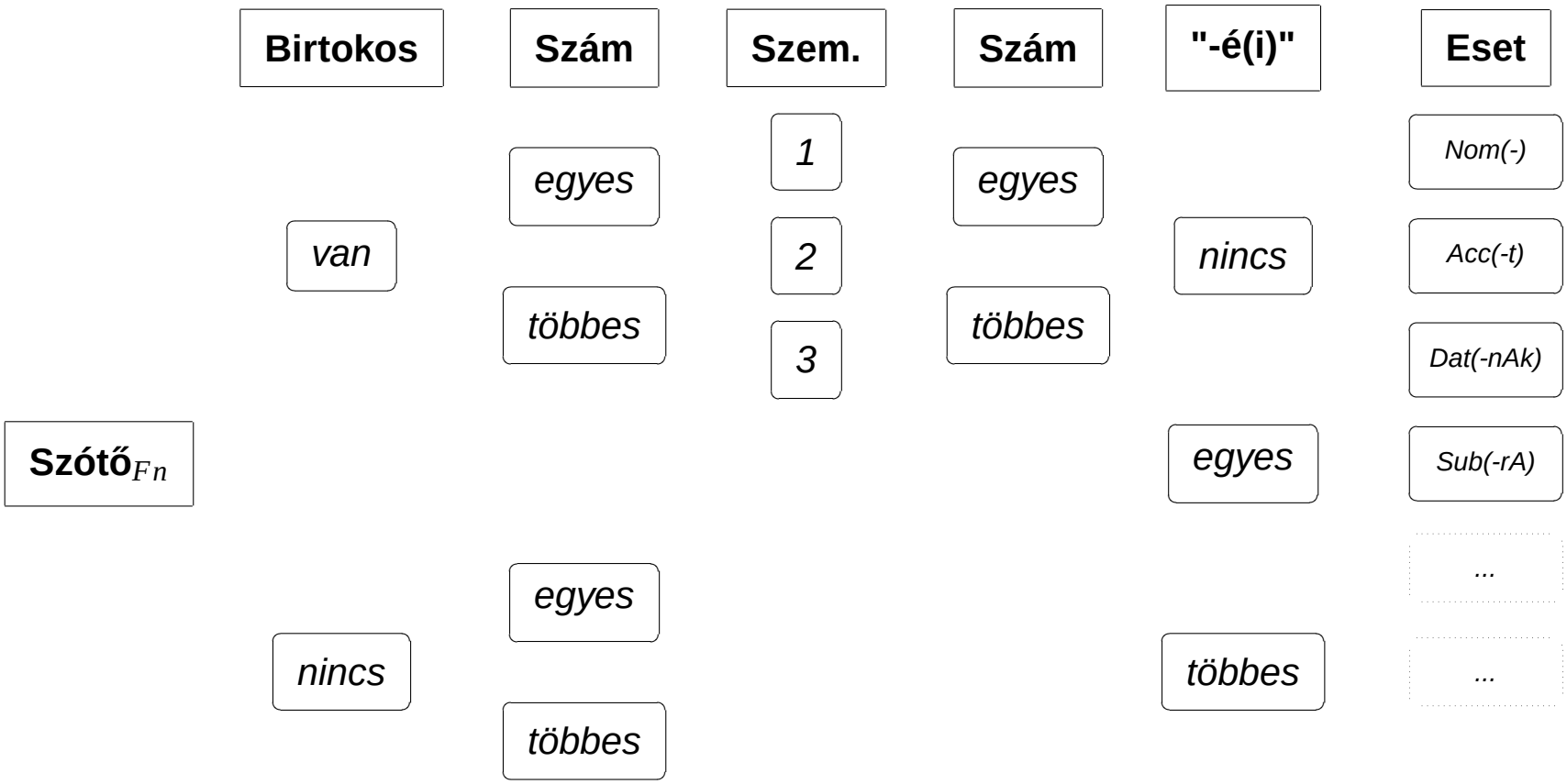
  Csak az veri fel az erdő csendjét, mivel az avar sír eddig feltáratlan maradt.

Számítógép és nyelvi elemzés

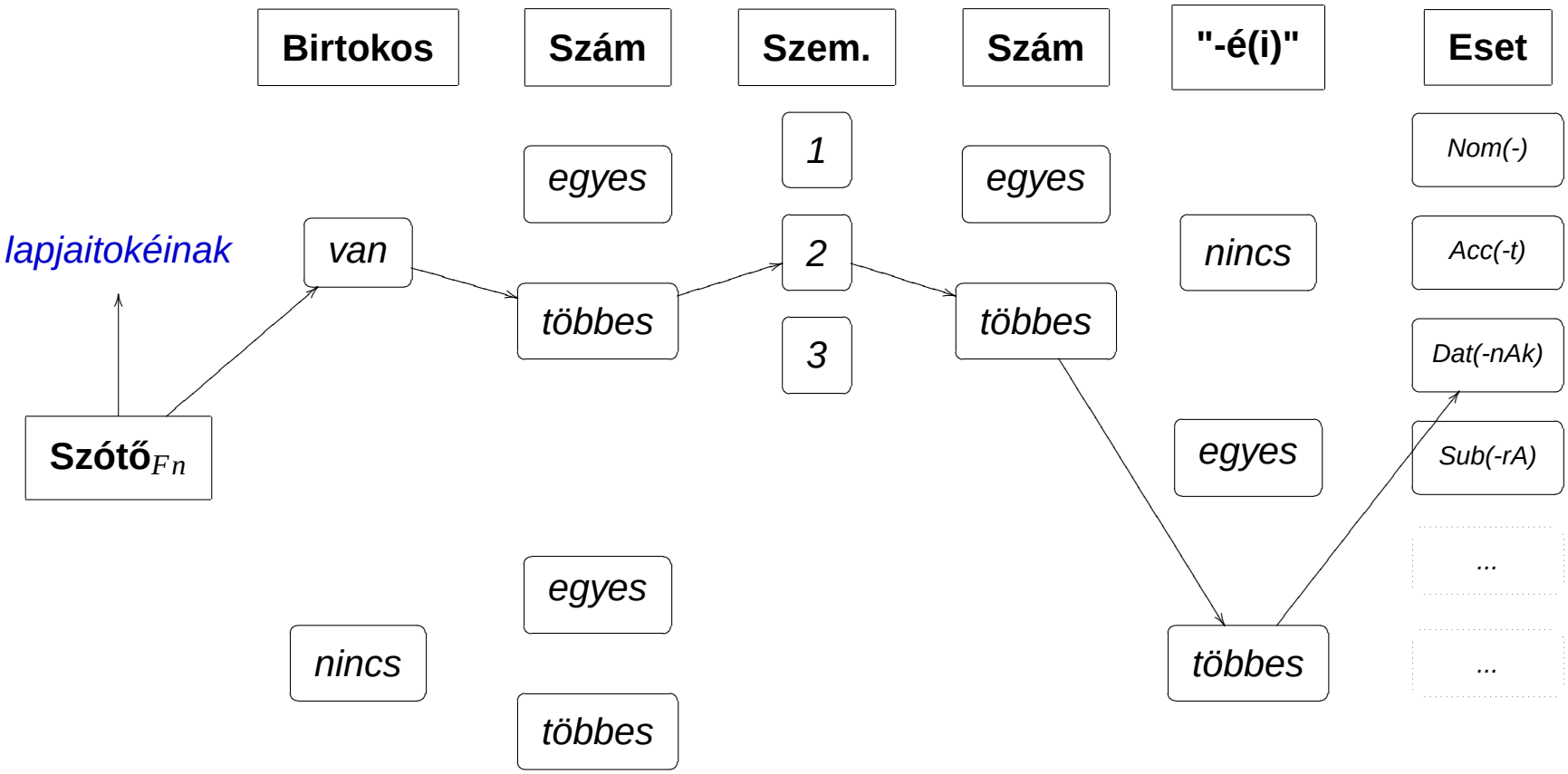
- homogén karakterfüzérből nyelvi jel(ek): határozzuk meg a kiterjedését és adjuk meg a tulajdonságait
- első lépésben a szóalakok mint elemi egységek szintjén
- *1 kódoló személy*
- MNSZ: 150 millió szó; 2 sec/szó (napi 24 órában) $\Rightarrow$ 9 év, 187 nap
- automatikus eljárás
- *morfoszintaktikai annotáció*
  - morfológiai elemzés
  - egyértelműsítés

Morfológiai elemzés – miért?

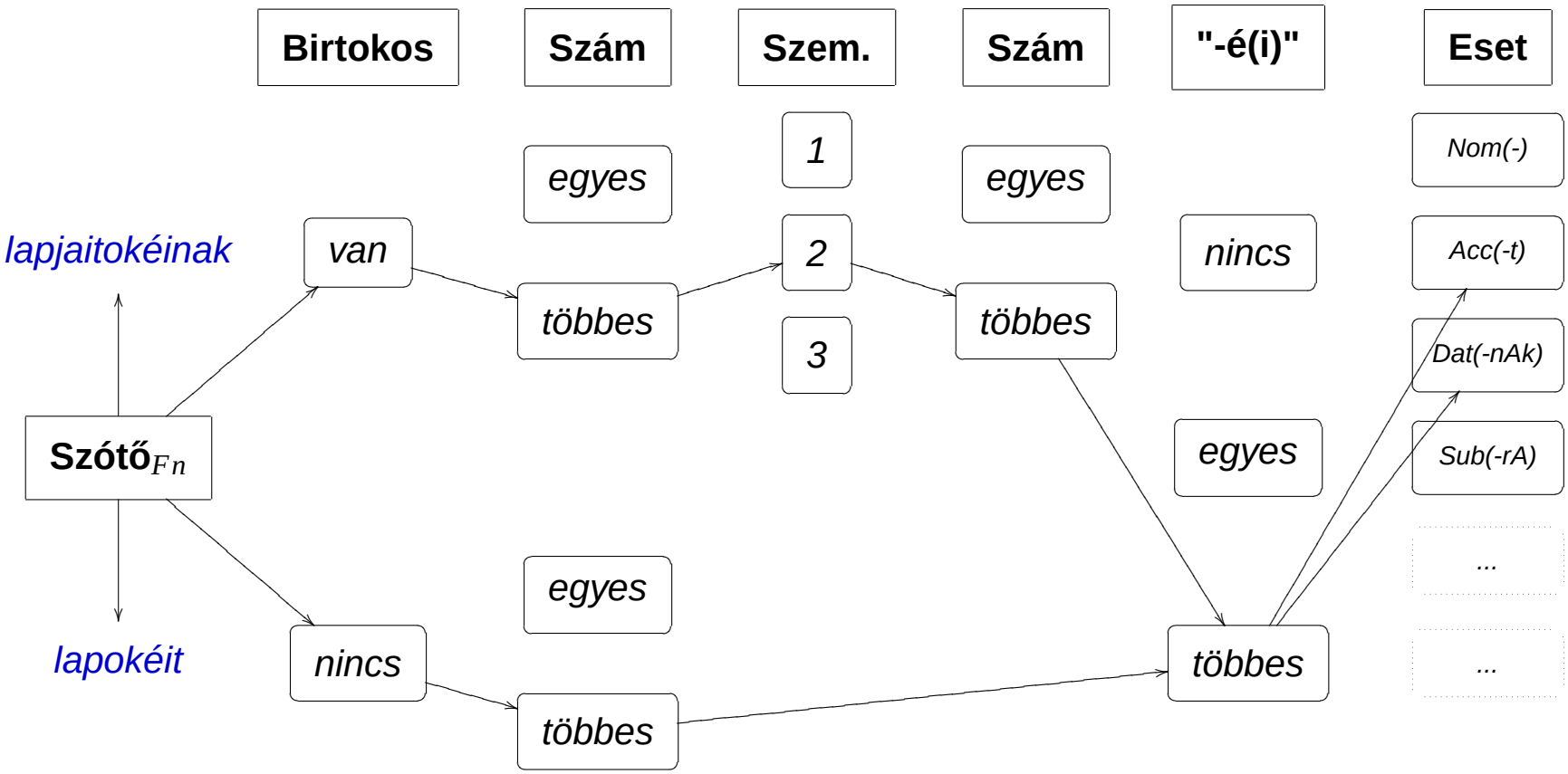
Morfológiai elemzés – miért?



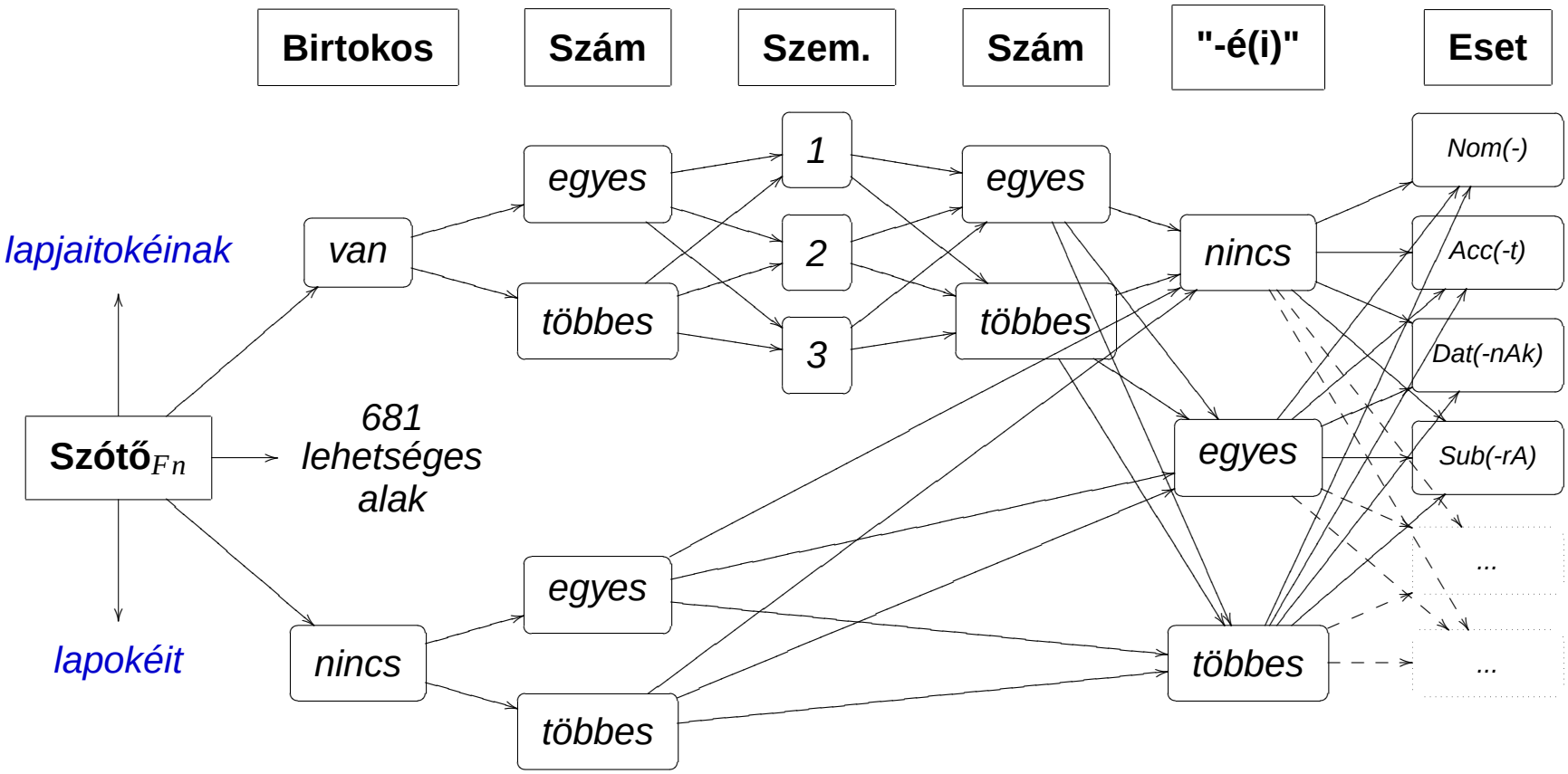
Morfológiai elemzés – miért?



Morfológiai elemzés – miért?



Morfológiai elemzés – miért?



A bemenő XML szöveg

Példa

```
<div type="test">
<head>Mire jó a nyelvtechnológia?</head>
<opener>
  <dateline>MTA
   <date iso8601="2003-11-04">2003. november 4.</date>
  </dateline>
</opener>
<p>
Az őszi avar sír a lába alatt.
Csak az veri fel az erdő csendjét,
mivel az avar sír eddig feltáratlan maradt.
</p>
</div>
```

Szegmentálás és morfológiai elemzés

- bemenő folyó szöveg mondatokra tagolása és a mondatok egyes szavakra bontása

- morfológiai elemző: karakterfüzérekhez mint szóalakokhoz hozzárendeli minden lehetséges morfológiai elemzésüket

Morfológiailag elemzett szöveg

Példa

```
...
4*1     TOK     MTA          BOS     MTA*[N][NOM]
5*1     DATE    2003._november_4.    EOS  2003._november_4.*[DATUM]
#               )SENT   </S>
#               (SENT   <S>
7*1     TOK     Az           BOS     az*[Det]|az*[Pro][NOM]
7*4     TOK     őszi                 őszi*[A][NOM]
7*9     TOK     avar                 avar*[A][NOM]|avar*[N][NOM]
7*14    TOK     sír                  sír*[N][NOM]|sír*[V][e3]
7*18    TOK     a                    a*[Det]
7*20    TOK     lába                 láb*[N][PSe3][NOM]
7*25    TOK     alatt                alatt*[Adv]|alatt*[NU]
7*30    PTERM   .            EOS     .*SPUNCT
#               )SENT   </S>
#               (SENT   <S>
```

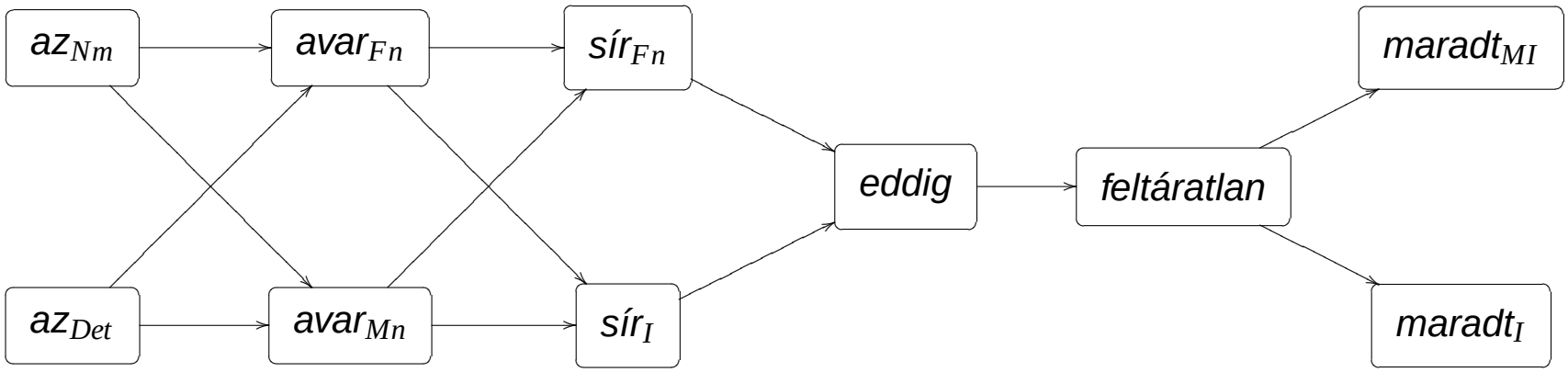
```
8*1     TOK     Csak     BOS     csak*[Adv]
8*6     TOK     az               az*[Det]|az*[Pro][NOM]
8*9     TOK     veri             ver*[V][Te3]
8*14    TOK     fel              fel*[Adv]|fel*[N][NOM]|fel*[Pre]
8*18    TOK     az               az*[Det]|az*[Pro][NOM]
8*21    TOK     erdő             erdő*[N][NOM]
8*26    TOK     csendjét          csend*[N][PSe3][ACC]
8*34    PUNCT   ,                ,*WPUNCT
9*1     TOK     mivel            mivel*[Con]|mi*[Pro][INS]|mivel*[Adv]
9*7     TOK     az               az*[Det]|az*[Pro][NOM]
9*10    TOK     avar             avar*[A][NOM]|avar*[N][NOM]
9*15    TOK     sír              sír*[N][NOM]|sír*[V][e3]
9*19    TOK     eddig            ez*[Pro][TER]
9*25    TOK     feltáratlan       feltáratlan*[A][NOM]
9*37    TOK     maradt           maradt*[MIB][NOM]|marad*[V][Me3]
9*43    PTERM   .        EOS     .*SPUNCT
#                )SENT    </S>
```

Többértelműség

- *kezdtek, végeztek, terem, állam, köröm, hullám, tanára, művére, női, sort, bájt, termet, nemzeti, feji, telefon, mondat, lejár, élek, sírok, laknak, falnak, halnak, telefonnak, váza, kacsa, héja, léptet, ereszt, béget, sikerül, települ, diák, torok, tubák, törtet, kopaszt, horpaszt, kisebben, fürgébben, adunk, kapunk, tudatunk*

Többértelműség

- *kezdtek, végeztek, terem, állam, köröm, hullám, tanára, művére, női, sort, bájt, termet, nemzeti, feji, telefon, mondat, lejár, élek, sírok, laknak, falnak, halnak, telefonnak, váza, kacsa, héja, léptet, ereszt, béget, sikerül, települ, diák, torok, tubák, törtet, kopaszt, horpaszt, kisebben, fürgébben, adunk, kapunk, tudatunk*



Morfoszintaktikai egyértelműsítés

- lehetséges elemzések közül a szövegkörnyezetbe, az adott mondatba illő kiválasztása
- *1 kódoló személy*
- MNSZ: 150 millió szó; kb. 23% többértelmű; 1 sec/szó (napi 24 órában) $\Rightarrow$ 1 év, 35 nap
- nagy mennyiségű, változatos típusú szöveg $\Rightarrow$ gyors, a változatosságot jól kezelő automatikus módszer
- relatív gyakoriságon alapuló eljárás: az egyes elemzések gyakoriságát, valamint (legfeljebb) szóhármasok elemzésének gyakoriságát veszi figyelembe (másodrendű rejtett Markov-modell)

Morfoszintaktikai egyértelműsítés

- a számítógépet meg kell tanítani a helyes elemzés kiválasztására
- 270 ezer szavas kézzel egyértelműsített tanító korpusz (17 óra) $\Rightarrow$ *nyelvi modell*
- adott kontextusban legvalószínűbb elemzés kiválasztása a nyelvi modellben tárolt információ alapján
- egyszerű modell: 97.5–98%-os teljesítmény

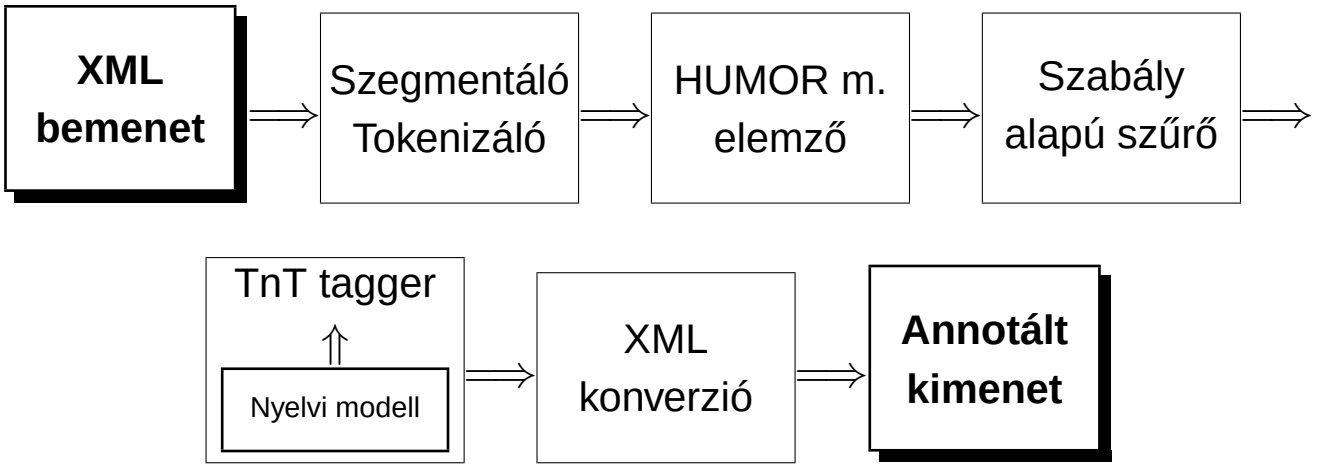
Szabály alapú modul

- egyértelműen megadható feltételek fennállása esetén 100%-os pontossággal működő szabályok
- 10%-kal csökkenthető a rosszul egyértelműsített esetek száma

```
2. 'az([Pro]|[Det])'
  - choose [Det] if followed by [N] beginning with vowel
  - choose [Pro] if followed by [Det] or [V] or [Con] or
    small case consonant or 'az'
? x.token=az x.msd={[Pro],[Det]}
+ [Det] f.msd=[N] f.bw=aáeéiíoóöőuúüűAÁEÉIÍOÓÖŐUÚÜŰ
+ [Pro] f.msd={[Det],[V],[Con]}
+ [Pro] f.bw=qwrtpsdfghjklmnbvcxz,:;.?!
+ [Pro] f.token=az
```

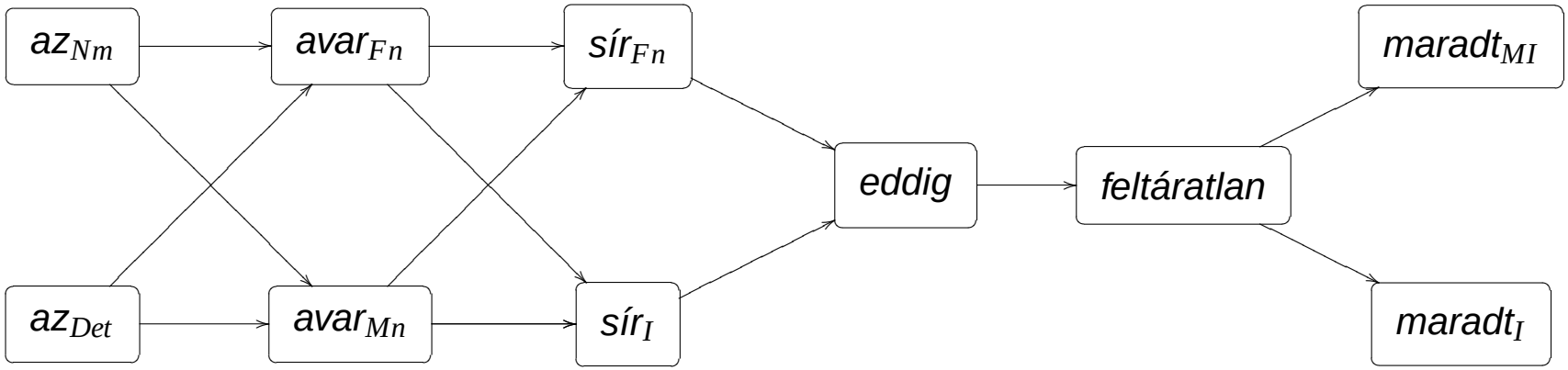
1. ábra. Egy egyértelműsítő szabály

Az egyértelműsítő eszközlánc

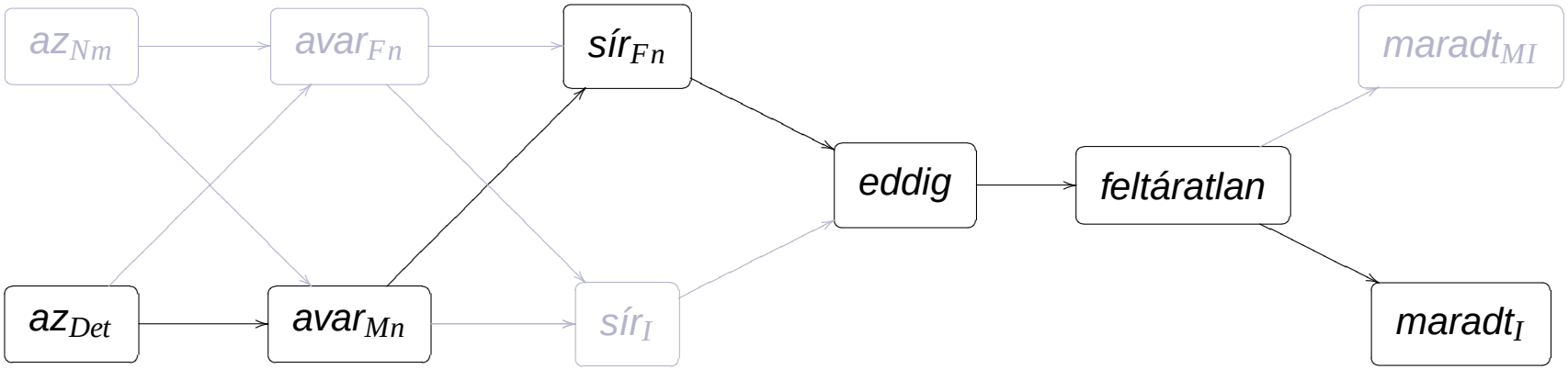


2. ábra. Az egyértelműsítő lánc komponensei

A kiválasztott elemzés



A kiválasztott elemzés



A végleges XML kimenet

Példa

```
<div type="test">
<head>
<s>
<w lemma="mire" msd="Adv">Mire</w>
<w lemma="jó" msd="A.NOM">jó</w>
<w lemma="a" msd="Det">a</w>
<w lemma="nyelvtechnológia" msd="N.NOM">nyelvtechnológia</w>
<c lemma="?" msd="SPUNCT">?</c>
</s>
</head>
<opener>
<dateline>
<w lemma="MTA" msd="N.NOM">MTA</w>
<date iso8601="2003-11-04">
```

```
<w lemma="2003._november_4." msd="DATUM">2003._november_4.</w>
</date>
</dateline>
</opener>
<p>
<s>
<w lemma="az" msd="Det">Az</w>
<w lemma="őszi" msd="A.NOM">őszi</w>
<w lemma="avar" msd="N.NOM">avar</w>
<w lemma="sír" msd="V.e3">sír</w>
<w lemma="a" msd="Det">a</w>
<w lemma="láb" msd="N.PSe3.NOM">lába</w>
<w lemma="alatt" msd="NU">alatt</w>
<c lemma="." msd="SPUNCT">.</c>
</s>
<s>
<w lemma="csak" msd="Adv">Csak</w>
<w lemma="az" msd="Pro.NOM">az</w>
```

```
<w lemma="ver" msd="V.Te3">veri</w>
<w lemma="fel" msd="Pre">fel</w>
<w lemma="az" msd="Det">az</w>
<w lemma="erdő" msd="N.NOM">erdő</w>
<w lemma="csend" msd="N.PSe3.ACC">csendjét</w>
<c lemma="," msd="WPUNCT">,</c>
<w lemma="mivel" msd="Adv">mivel</w>
<w lemma="az" msd="Det">az</w>
<w lemma="avar" msd="A.NOM">avar</w>
<w lemma="sír" msd="N.NOM">sír</w>
<w lemma="ez" msd="Pro.TER">eddig</w>
<w lemma="feltáratlan" msd="A.NOM">feltáratlan</w>
<w lemma="marad" msd="V.Me3">maradt</w>
<c lemma="." msd="SPUNCT">.</c>
</s>
</p>
</div>
```

Összefoglalás

- már a gépi nyelvfeldolgozás kezdetén is számos olyan feladatot kell megoldani, ami a beszélők számára triviális
- megkerülhetetlen lépések minden további nyelvfeldolgozó alkalmazás számára
- a bemutatott eljárás gyakorlati alkalmazása: MNSZ egyértelműsítése

VÉGE