

Magyar Tudományos Akadémia

**Nyelvtudományi Intézet**

# **Egy szónak is száz a vége**

Oravecz Csaba

MTA Nyelvtudományi Intézet  
Korpusznyelvészeti osztály  
oravecz@nytud.hu

## Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

karaktersorozatokat

- $|a|z|_|a|v|a|r|_|s|í|r|$ : 11 azonos típusú elemi egység

## Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

karaktersorozatokat

- $|a|z|_|a|v|a|r|_|s|i|r|$ : 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező nyelvi jelekké állnak össze

## Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

karaktersorozatok

- $|a|z|_|a|v|a|r|_|s|í|r|$ : 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező nyelvi jelekké állnak össze
- a környezetüktől függően alapvető tulajdonságaik változtathatnak

## Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

karaktorsorozatok

- $|a|z|_|a|v|a|r|_|s|í|r|$ : 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező nyelvi jelekké állnak össze
- a környezetüktől függően alapvető tulajdonságaik változtathatnak

Az őszi avar sír a lába alatt.

## Bevezetés

- mit lát a számítógép a természetes nyelvi megnyilatkozásokból?

karaktorsorozatok

- $|a|z|_|a|v|a|r|_|s|í|r|$ : 11 azonos típusú elemi egység
- a beszélők számára azonban számos fontos tulajdonsággal rendelkező nyelvi jelekké állnak össze
- a környezetüktől függően alapvető tulajdonságaik változtathatnak

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

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

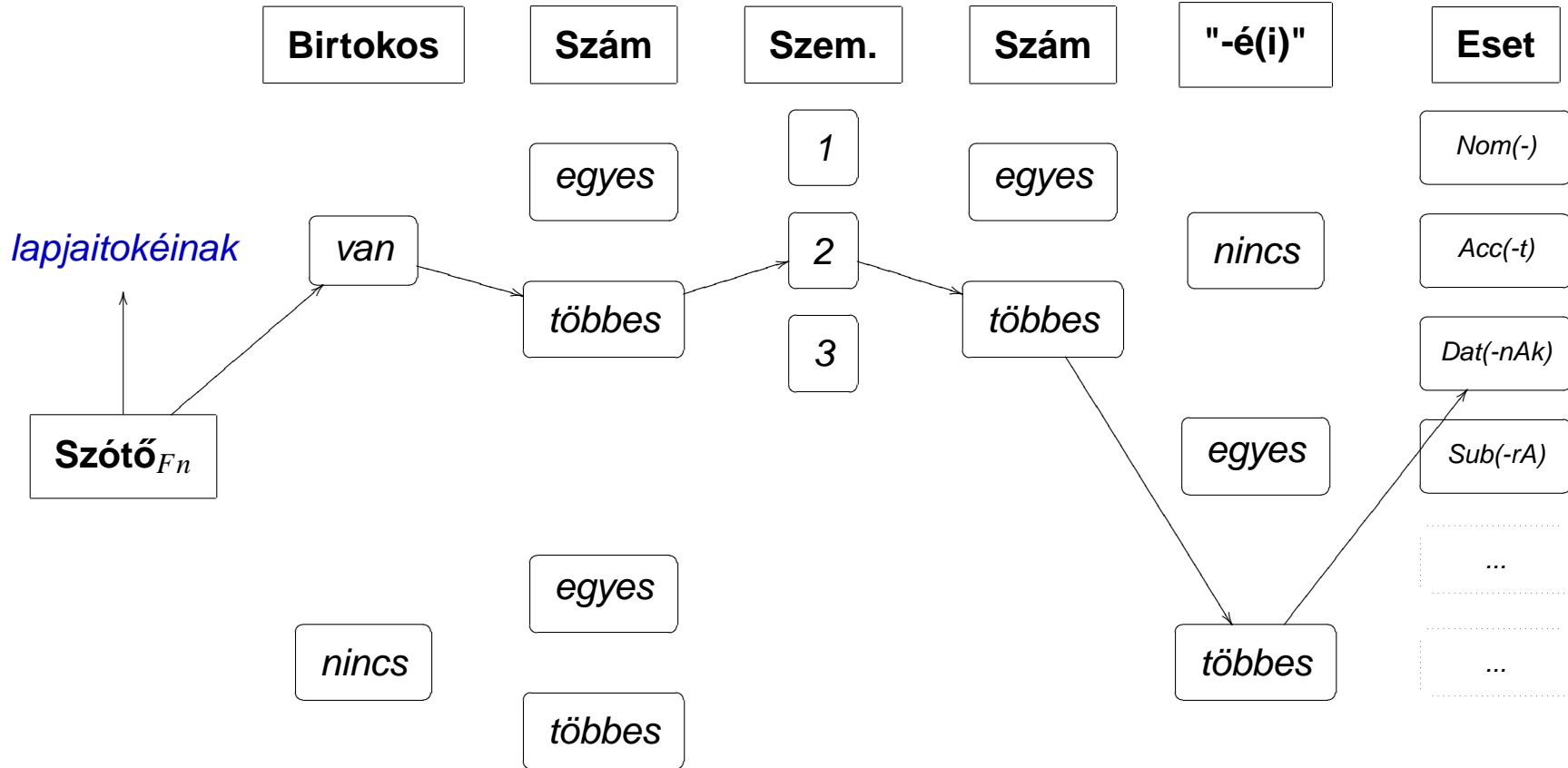
# Morfológiai elemzés – miért?



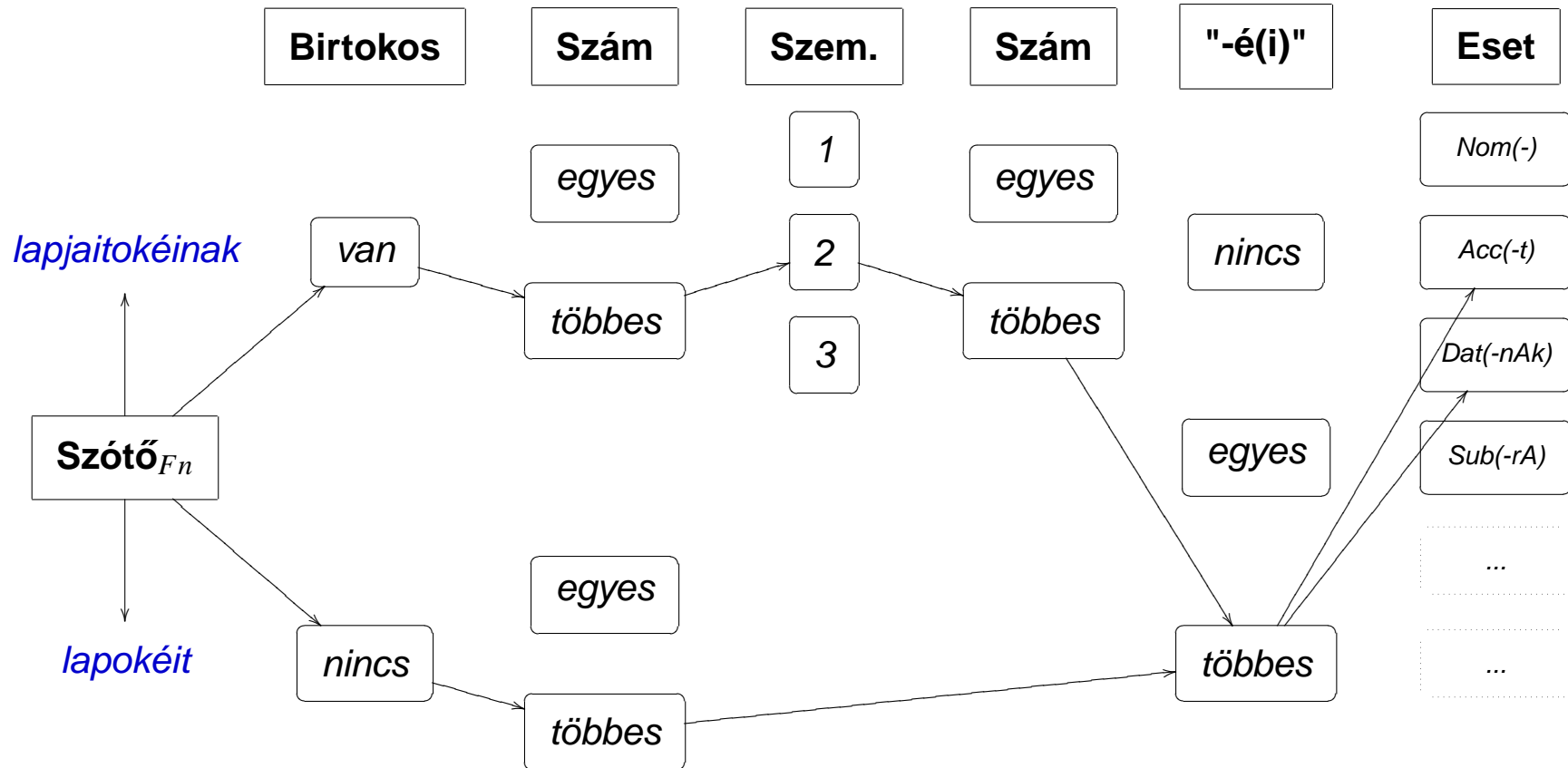
## Morfológiai elemzés – miért?

	Birtokos	Szám	Szem.	Szám	"-é(i)"	Eset
		egy	1	egy		Nom(-)
	van	több	2	több	nincs	Acc(-t)
			3			Dat(-nAk)
Szótő <sub>Fn</sub>		egy			egy	Sub(-rA)
	nincs	több			több	...
						...

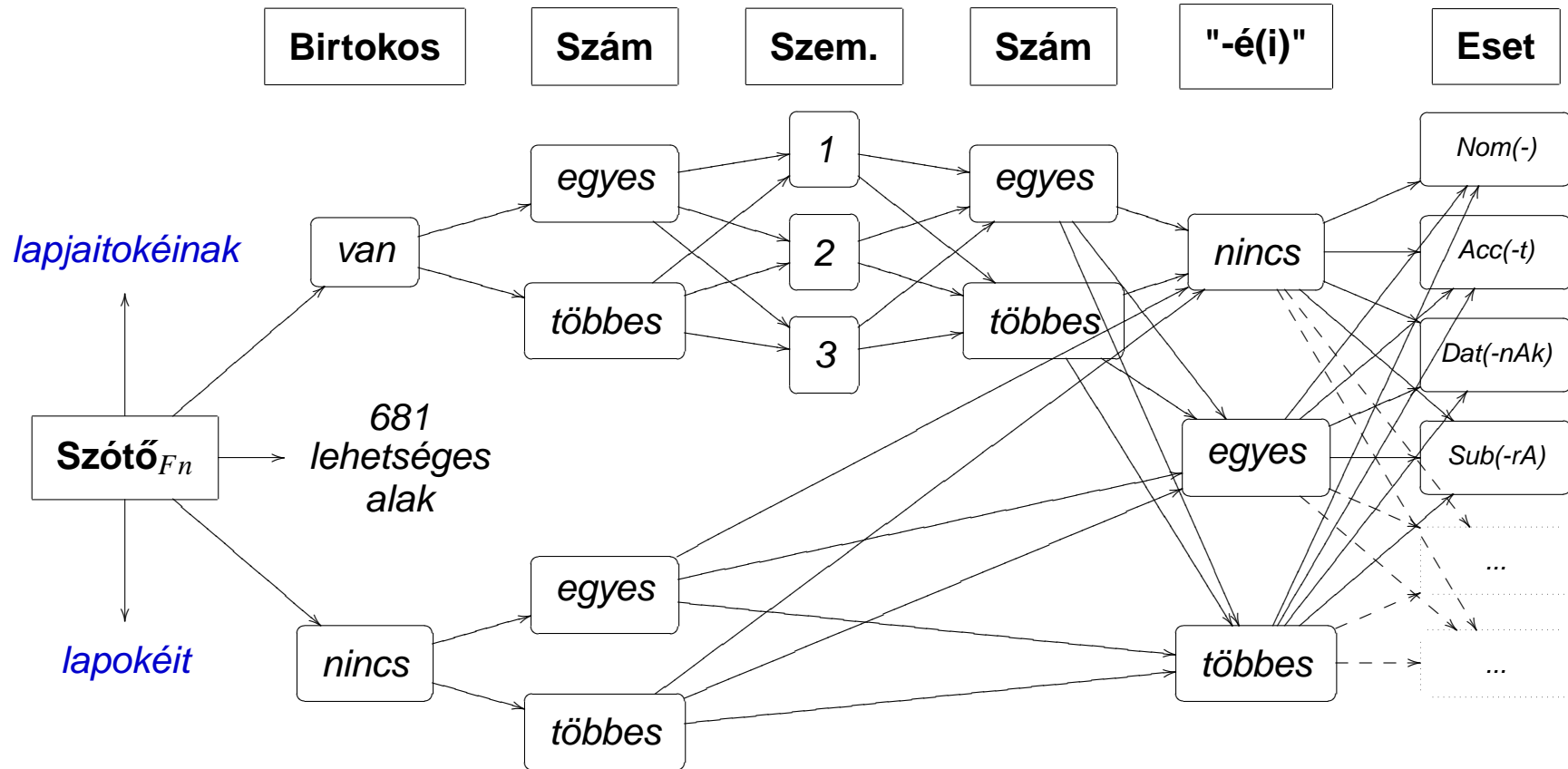
# Morfológiai elemzés – miért?



# Morfológiai elemzés – miért?



# Morfológiai elemzés – miért?



## A bemenő XML szöveg

Példa

```
<div type="test">
<head>Mire jó a nyelvtechnológia?</head>
<opener>
  <dateline>INFOtrend
    <date iso8601="2003-11-27">2003. november 27.</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: karakterfüzésekhez mint szóalakokhoz hozzárendeli minden lehetséges morfológiai elemzésüket

## Morfológiailag elemzett szöveg

Példa

```

...
4*1   TOK   INFOtrend   BOS   INFOtrend* [N] [NOM]
5*1   DATE  2003._november_27.   EOS   2003._november_27.* [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	<b>Csak</b>	BOS	csak* [Adv]
8*6	TOK	<b>az</b>		az* [Det]   az* [Pro] [NOM]
8*9	TOK	<b>veri</b>		ver* [V] [Te3]
8*14	TOK	<b>fel</b>		fel* [Adv]   fel* [N] [NOM]   fel* [Pre]
8*18	TOK	<b>az</b>		az* [Det]   az* [Pro] [NOM]
8*21	TOK	<b>erdő</b>		erdő* [N] [NOM]
8*26	TOK	<b>csendjét</b>		csend* [N] [PSe3] [ACC]
8*34	PUNCT	<b>,</b>		, *WPUNCT
9*1	TOK	<b>mivel</b>		mivel* [Con]   mi* [Pro] [INS]   mivel* [Adv]
9*7	TOK	<b>az</b>		az* [Det]   az* [Pro] [NOM]
9*10	TOK	<b>avar</b>		avar* [A] [NOM]   avar* [N] [NOM]
9*15	TOK	<b>sír</b>		sír* [N] [NOM]   sír* [V] [e3]
9*19	TOK	<b>eddig</b>		ez* [Pro] [TER]
9*25	TOK	<b>feltáratlan</b>		feltáratlan* [A] [NOM]
9*37	TOK	<b>maradt</b>		maradt* [MIB] [NOM]   marad* [V] [Me3]
9*43	PTERM	<b>.</b>	EOS	. *SPUNCT
#		) SENT	</S>	

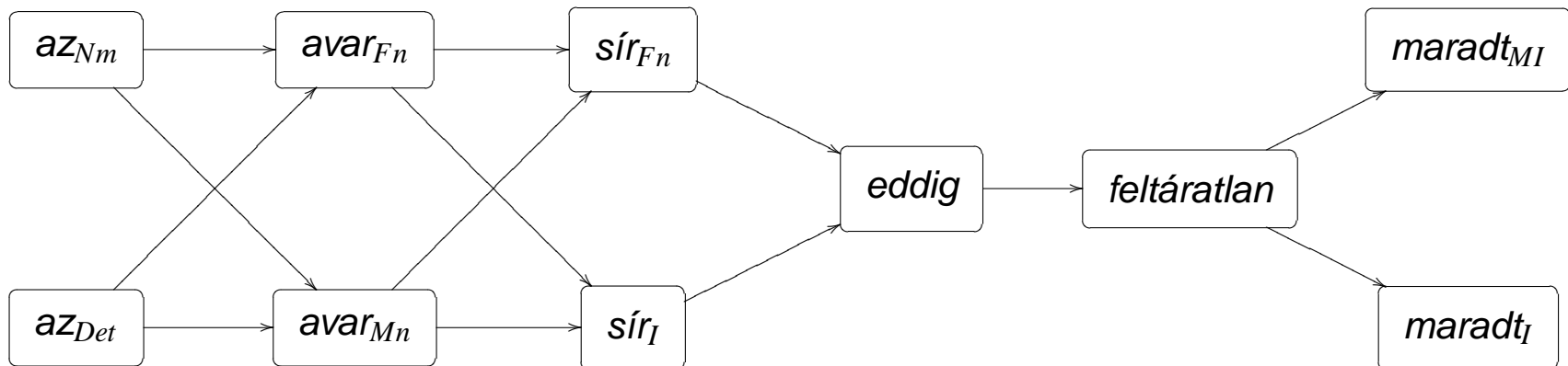


## **Alaktani többértelműség**

- *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, kisebbben, függébben, adunk, kapunk, tudatunk*

## Alaktani 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, függébben, adunk, kapunk, tudatunk*



## Morfoszintaktikai egyértelműsítés

- lehetséges elemzések közül a szövegekörnyezetbe, 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) szóhármások 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

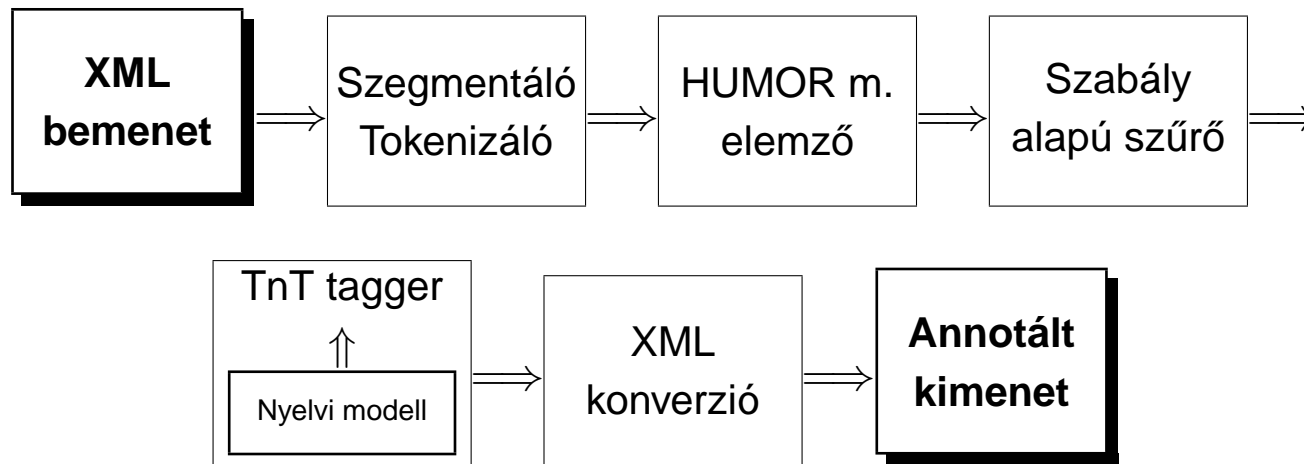
```

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.msds={ [Pro],[Det]}
+ [Det] f.msds=[N] f.bw=aáeéiíoóöőuúüüAÁEÉIÍOÓÖŐUÚÜÜ
+ [Pro] f.msds={ [Det],[V],[Con]}
+ [Pro] f.bw=qwrtpsdfghjklmnbvcxz,;:?!
+ [Pro] f.token=az

```

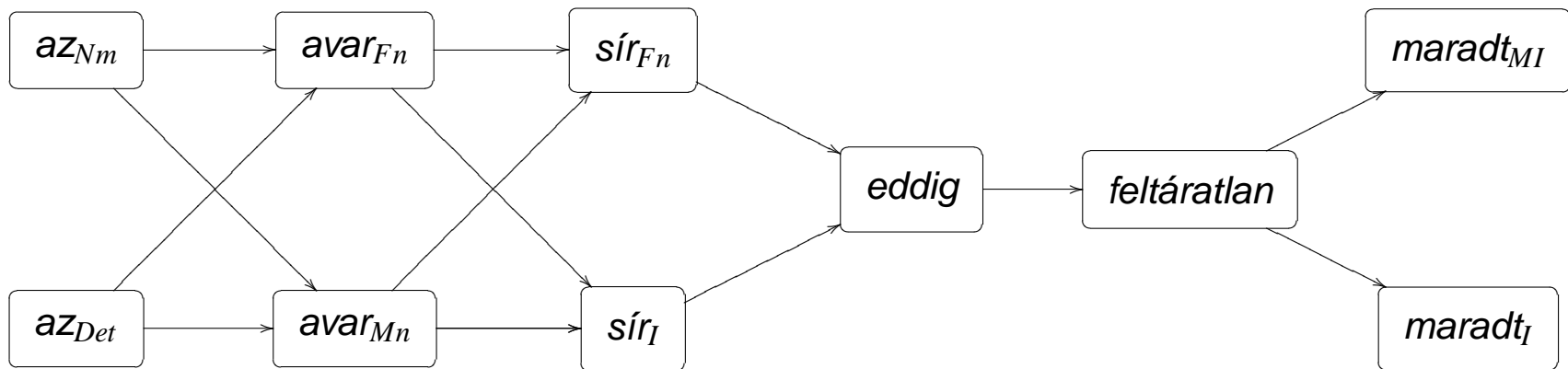
1. ábra. Egy egyértelműsítő szabály

## Az egyértelműsítő eszközlánc

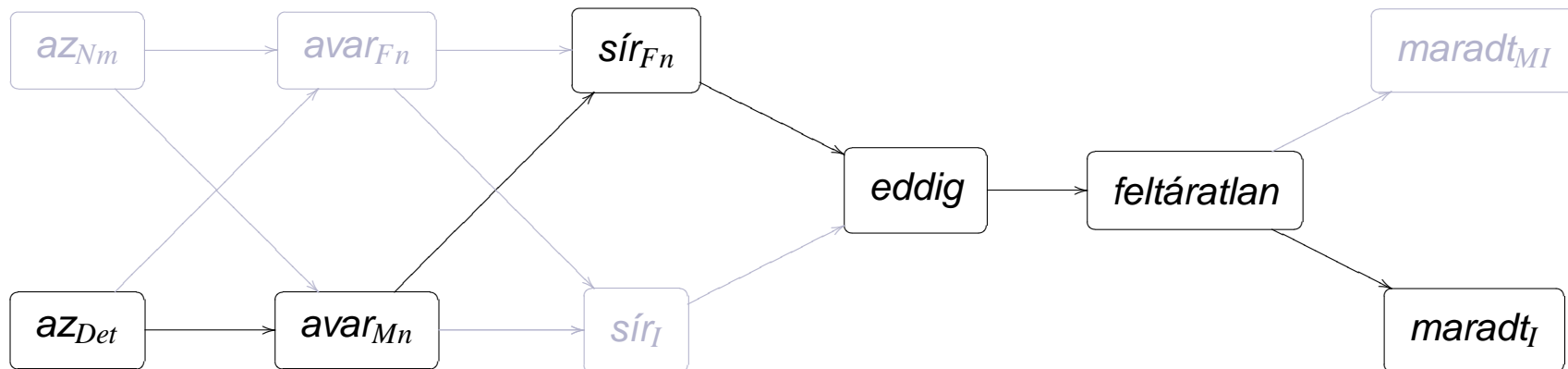


2. ábra. Az egyértelműsítő lánc komponensei

## 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="INFOtrend" msd="N.NOM">INFOtrend</w>
<date iso8601="2003-11-27">
```

```
<w lemma="2003._november_27." msd="DATUM">2003._november_27.</w>
</date>
</dateline>
</opener>
<p>
<s>
<w lemma="az" msd="Det">Az</w>
<w lemma="ősz" msd="A.NOM">Ősz</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 többértelműség a nyelvfeldolgozás minden szintjén megoldandó problémát jelent
- a bemutatott eljárás gyakorlati alkalmazása: MNSZ egyértelműsítése

VÉGE