Nyelvelemzés sajátkezűleg a magyar INTEX rendszer

Váradi Tamás
varadi@nytud.hu

Vázlat

- A történet eddig
- Az INTEX rendszer
- A magyar modul
- Az INTEX korlátai
- premier előtt: NooJ
- konklúziók, további teendők

Rövid történeti háttér

- LADL – Maurice Gross
  - lexikai nyelvtan, véges állapotú technológia
- 1993 INTEX **Max Silberztein**
- elterjedtség
  - több mint 20 kutatóintézetben alkalmazzák
  - francia, angol, szerb, portugál,spanyol
  - éves konferenciák
- Magyar?

  Premier – itt és most!

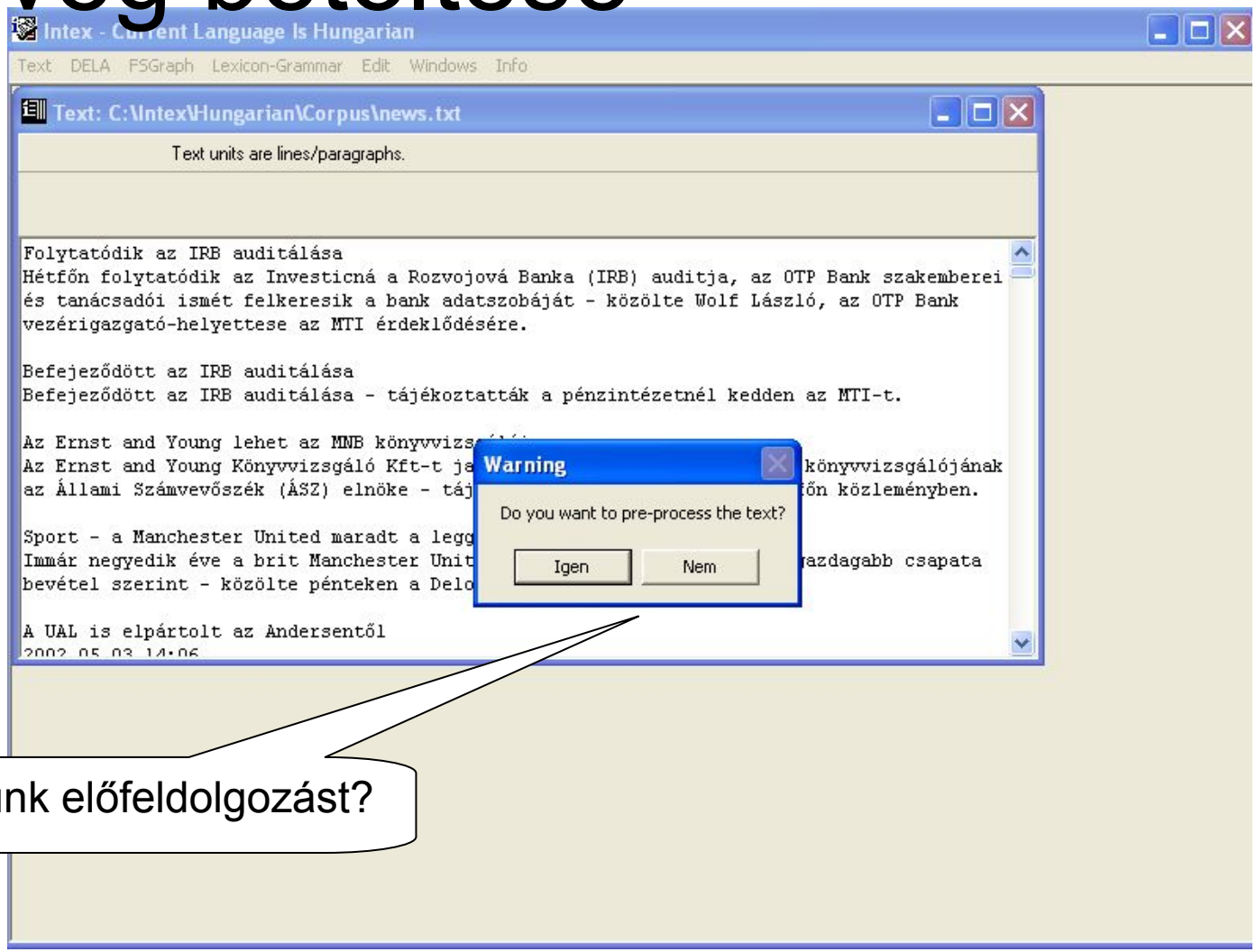
Integrált nyelvelemző környezet

- Három eszköz egyben
  - korpuszelemző
  - nyelvtanfejlesztő
  - oktató
- Egységes (véges állapotú) technológia
- Robosztus, jól kidolgozott lexikon
- morfoszintaktikai, szemantikai jegyek
- Grafikus felület

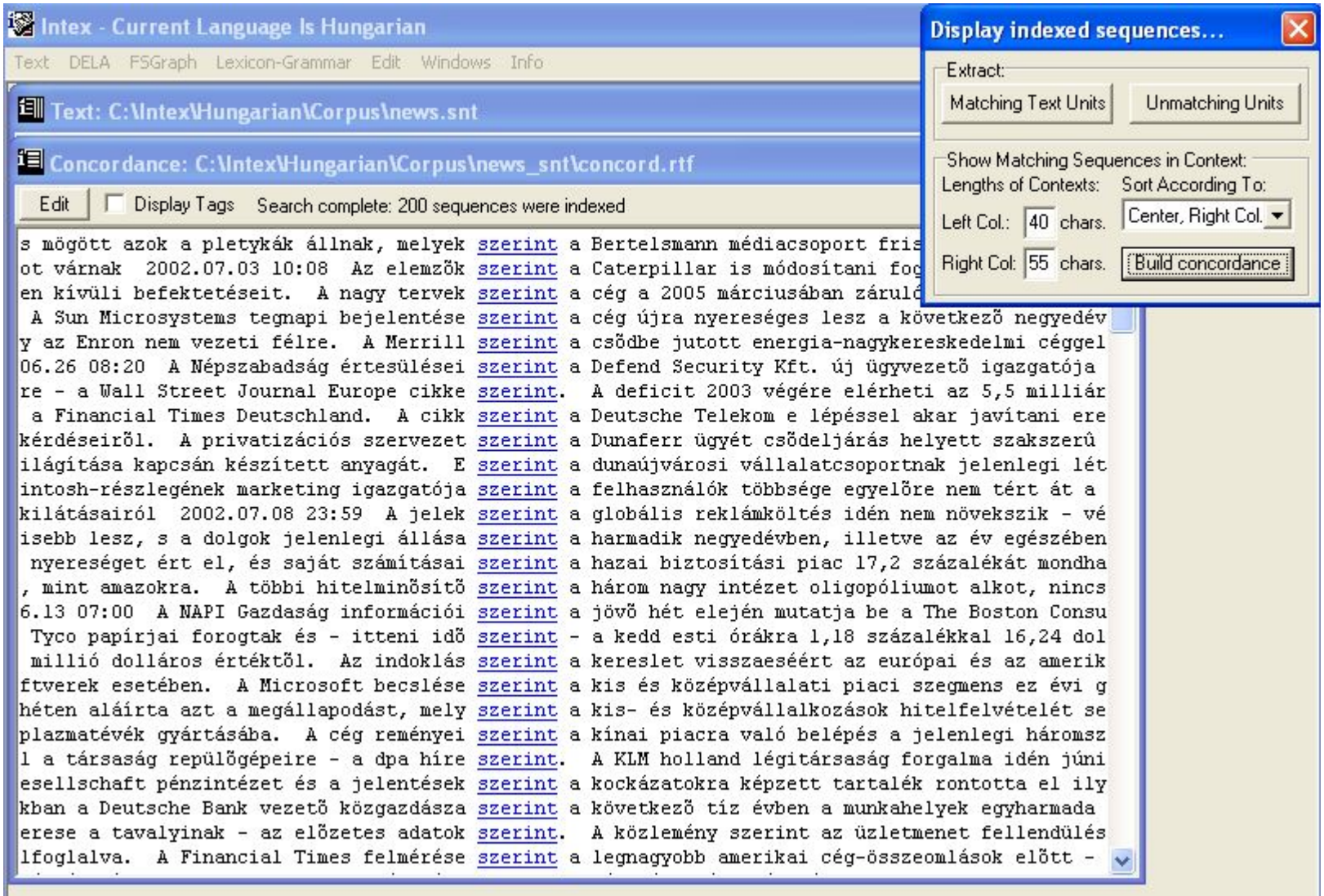
INTEX a korpuszkezelő eszköz

- **.txt** szövegfile → azonnali lekérdezés
- keresés
  - reguláris kifejezések
  - gráfok (nyelvtanok)
- kimenet: konkordancia

A szöveg betöltése



egyszerű szöveglekérdezés

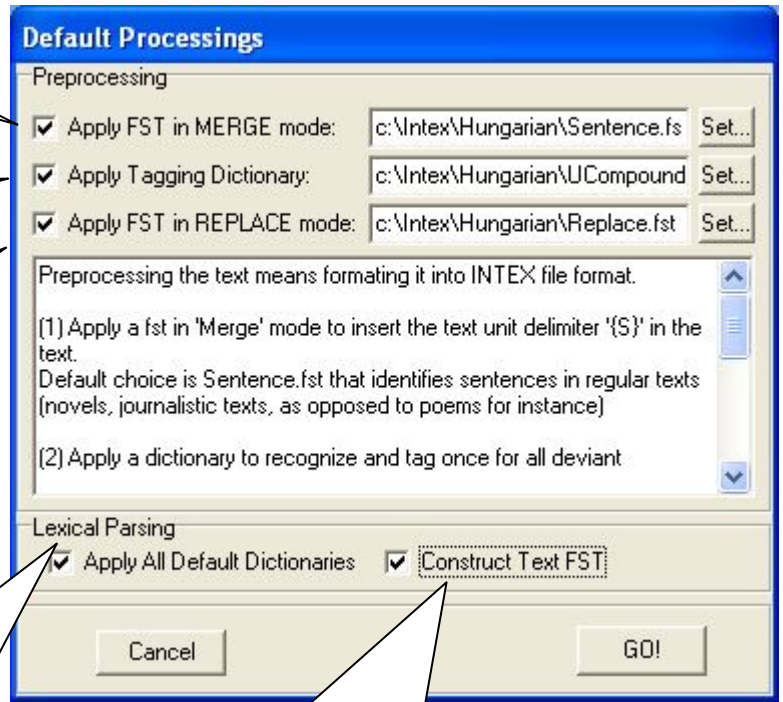


A szöveg előfeldolgozása

Mondatokra bontás
{S} címke beillesztésével

Többszavas
kifejezések bejelölése

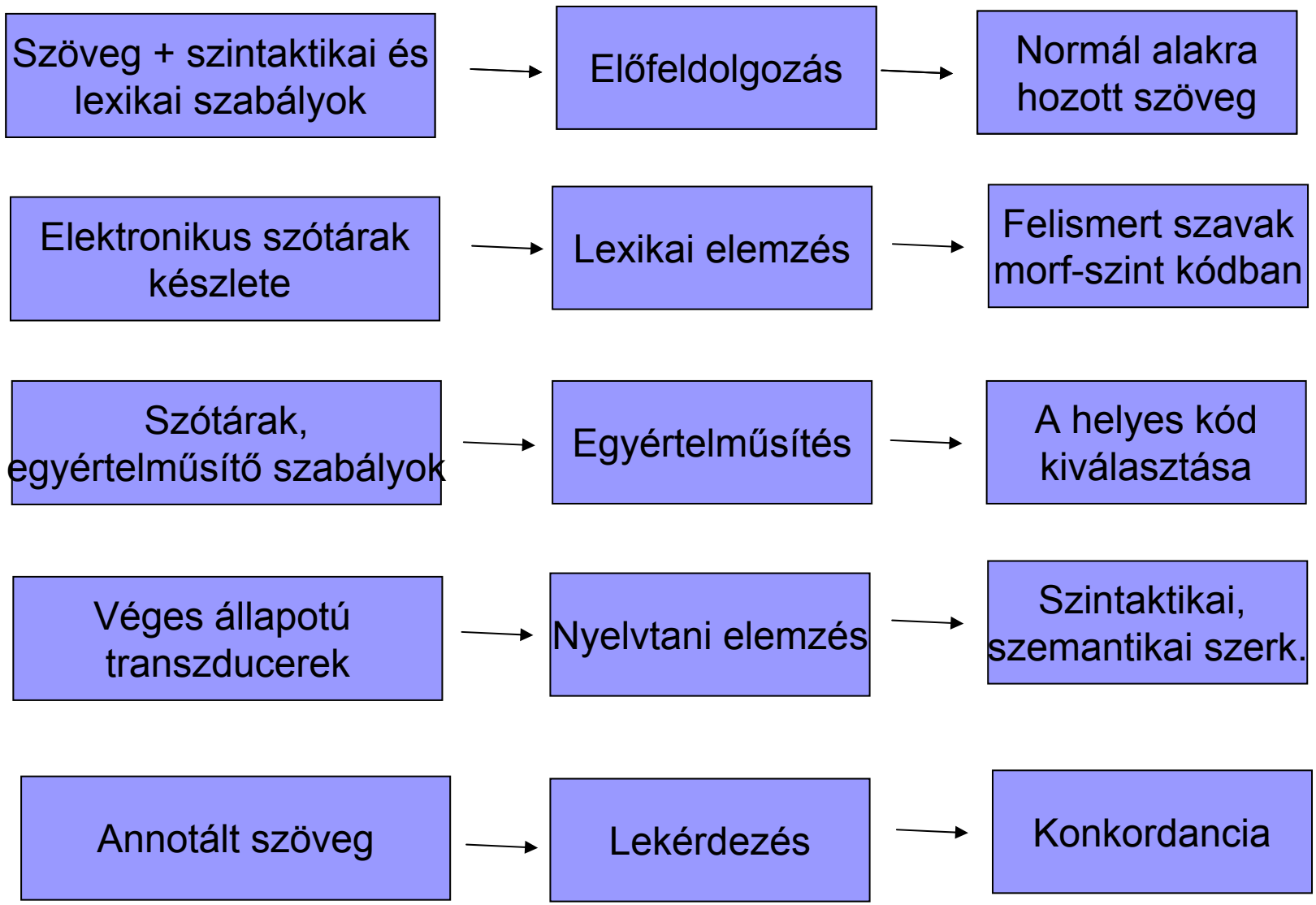
Nem „szabványos” alakok
standard alakra konvertálása
(aszem-> azt hiszem)



Az alapszótárak lefuttatása

Mondatok véges állapotú
automatákba szervezése

A korpusz teljes feldolgozása



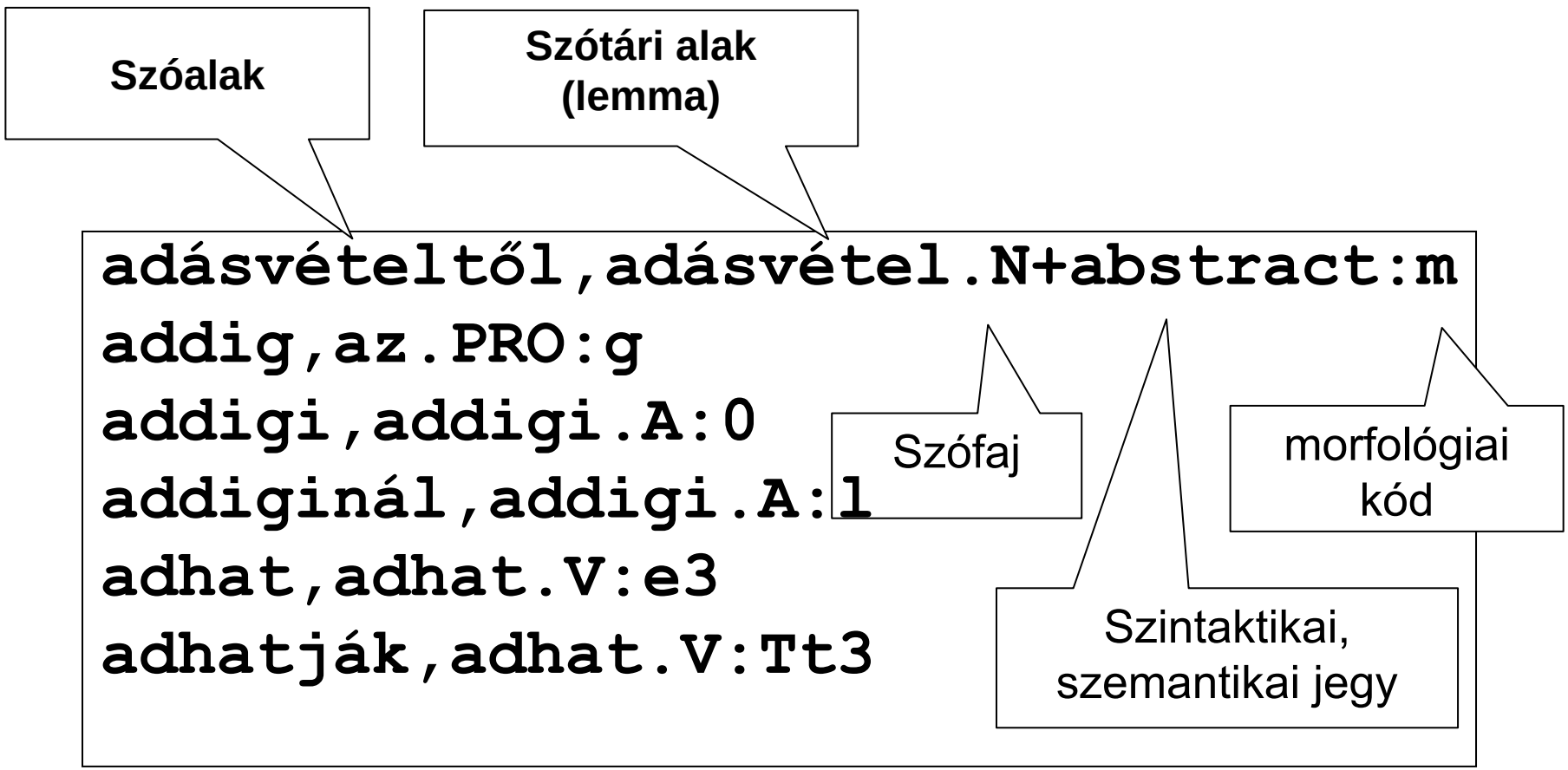
Lexikai erőforrások

- szótárak
  - véges számú, lexikalizált elemek
  - morfológiai, szintaktikai, szemantikai jegyek
- lexikai szabályok (lexikai transzducerek)
  - hagyományos szótárak által ignorált elemek
    - nyitott lexikai osztályok (pl. képzett elemek)
    - számok, dátumok, nevek (nyílt tokenosztályok)

Az INTEX szótárak

- központi szerep
- robosztus, gyors,
- nyitott, bővíthető
- 4 féle szótár
  - DELAF egyelemű szóalakok
  - DELACF többszavas kifejezések(szóalakok)
  - DELAE idiomatikus kifejezések
  - DELAS morfológiai generáló szótár

A DELAF szótár



A DELACF szótár

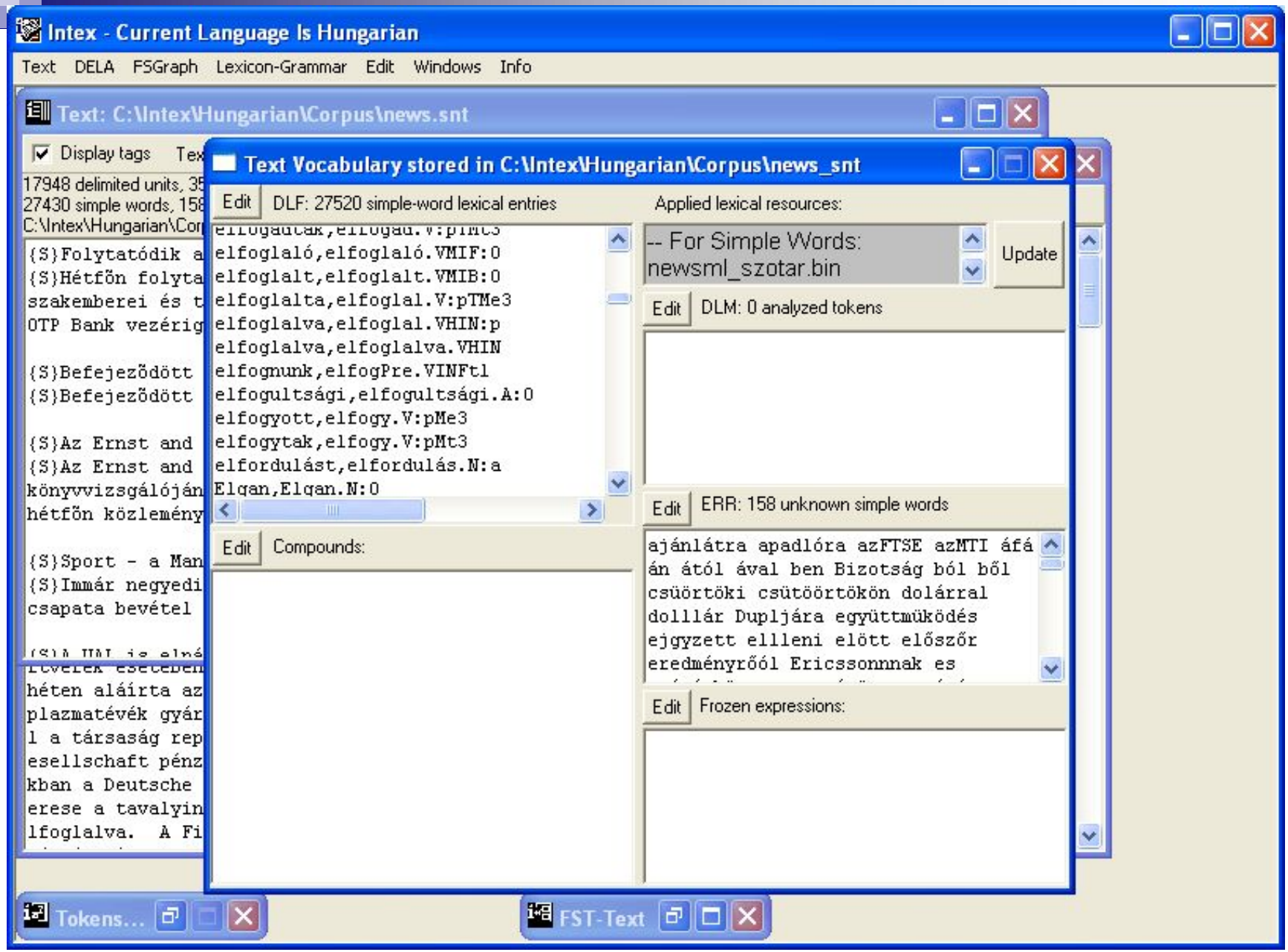
Egye
ezze
talá
Belü

```
Bayer Corp,.N+proper:0
Bayerische Hypo,.N+proper:0
Beck György,.N+proper:0
Befektetési Rt,.N+proper:0
Bellis Rt,.N+proper:0
Bencze József,.N+proper:0
Beogradska Banka,.N+proper:0
Bernard Ebbers,.N+proper:0
Berva Rt,.N+proper:0
Best Buy,.N+proper:0
Best Buy,.N+proper:0
BetonŁt Rt,.N+proper:0
Betonút Rt,.N+proper:0
Betonútépítő Rt,.N+proper:0
Bicinium Kft,.N+proper:0
Bild Zeitung,.N+proper:0
Bill Gates,.N+proper:0
Binder Rt,.N+proper:0
Biodiszkont Rt,.N+proper:0
Biologicals Kft,.N+proper:0
Body Shop,.N+proper:0
Boeing Co.,.N+proper:0
```

Ln 1 : 1 283 Col 1 Sel 0 | 33,32 KB | ANSI | CR+LF | INS | Default Text

A lexikai kódolás

- Az alapszótár toldalékolt szóalakok listája
- Több százezer szavas szótárak
- Tetszőleges információ kódolható szint./szem. jegy formájában
- Morfológai jegy kódja csak egy betű lehet
- Célszerű igazodni a kialakult gyakorlathoz
- A szótárak láncba szervezhetők és ki-be kapcsolhatók


Intex - Current Language Is Hungarian
Text DELA FSGraph Lexicon-Grammar Edit Windows Info
Text: C:\Intex\Hungarian\Corpus\news.snt
Display tags
17948 delimited units, 35
27430 simple words, 158
C:\Intex\Hungarian\Cor
{S}Folytatódik a
{S}Hétfõn folyta
szakemberei és t
OTP Bank vezérig
{S}Befejezõdött
{S}Befejezõdött
{S}Az Ernst and
{S}Az Ernst and
könyvvizsgálóján
hétfõn közlemény
{S}Sport - a Man
{S}Immár negyedi
csapata bevétel
héten aláírta az
plazmatévék gyár
l a társaság rep
esellschaft pénz
kban a Deutsche
erese a tavalyin
lfoglalva. A Fi
Text Vocabulary stored in C:\Intex\Hungarian\Corpus\news_snt
Edit
DLF: 27520 simple-word lexical entries
elfoglaló,elfoglaló.VMIF:0
elfoglalt,elfoglalt.VMIB:0
elfoglalta,elfoglal.V:pTMe3
elfoglalva,elfoglal.VHIN:p
elfoglalva,elfoglalva.VHIN
elfognunk,elfogPre.VINFtl
elfogultsági,elfogultsági.A:0
elfogyott,elfogy.V:pMe3
elfogytak,elfogy.V:pMt3
elfordulást,elfordulás.N:a
Elgan,Elgan.N:0
Edit
Compounds:
Applied lexical resources:
-- For Simple Words:
newsml_szotar.bin
Update
Edit
DLM: 0 analyzed tokens
Edit
ERR: 158 unknown simple words
ajánlátra apadlóra azFTSE azMTI áfá
án ától ával ben Bizotság ból ből
csüörtöki csütöörtökön dolárral
dolllár Dupljára együttmüködés
ejgyzett ellleni elött elõször
eredményrõól Ericssonnnak es
Edit
Frozen expressions:
Tokens...
FST-Text

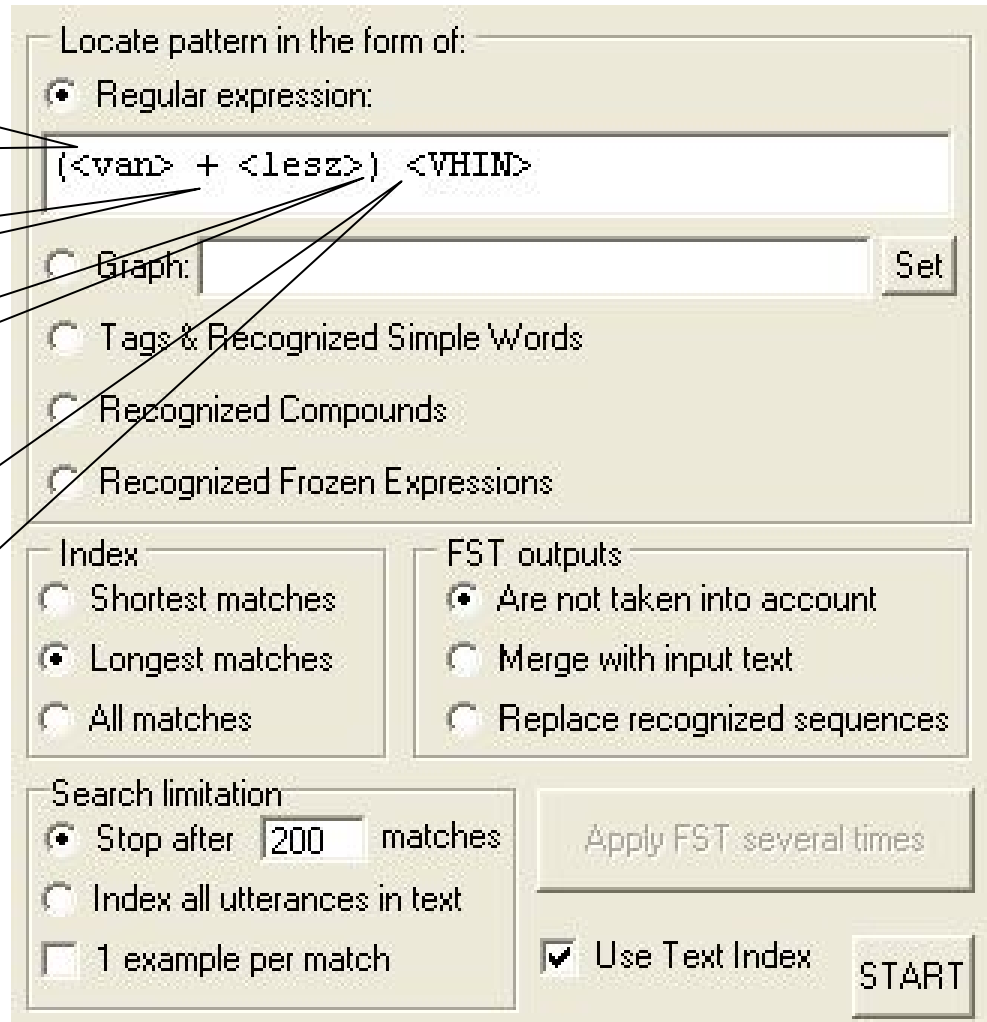
Lekérdezés képlettel

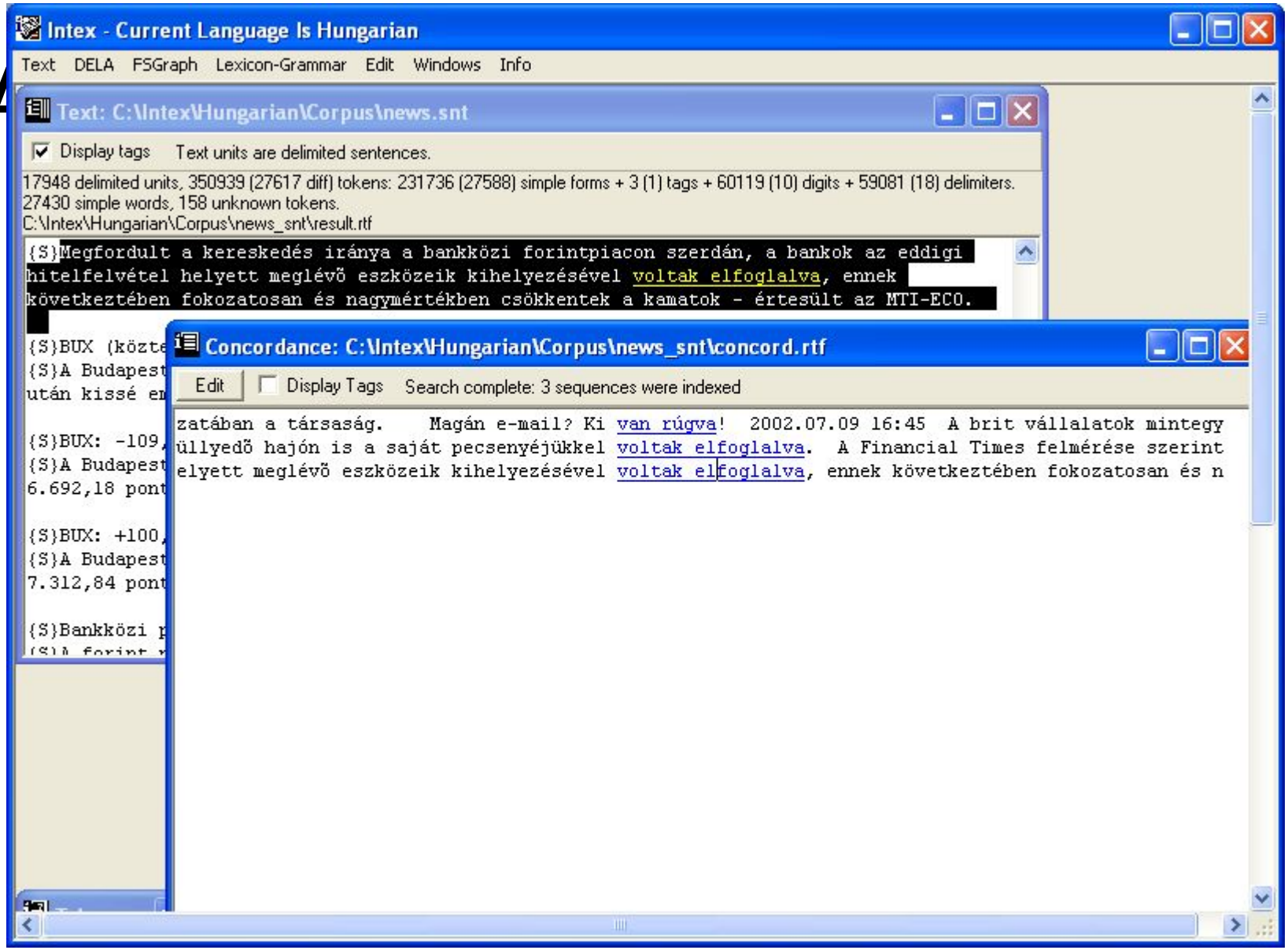
<van> = lemma, illeszkedik a van összes ragozott alakjára

\+ = a logikai „vagy” művelet jele

a zárójelben szereplő elemek közül az egyik

Szóköz = konkatenáció




Intex - Current Language Is Hungarian
Text DELA FSGraph Lexicon-Grammar Edit Windows Info
Text: C:\Intex\Hungarian\Corpus\news.snt
Display tags Text units are delimited sentences.
17948 delimited units, 350939 (27617 diff) tokens: 231736 (27588) simple forms + 3 (1) tags + 60119 (10) digits + 59081 (18) delimiters.
27430 simple words, 158 unknown tokens.
C:\Intex\Hungarian\Corpus\news_snt\result.rtf
{S}Megfordult a kereskedés iránya a bankközi forintpiacon szerdán, a bankok az eddigi hitelfelvétel helyett meglévő eszközeik kihelyezésével voltak elfoglalva, ennek következtében fokozatosan és nagymértékben csökkentek a kamatok - értesült az MTI-ECO.
Concordance: C:\Intex\Hungarian\Corpus\news_snt\concord.rtf
Edit Display Tags Search complete: 3 sequences were indexed
zatában a társaság. Magán e-mail? Ki van rúgva! 2002.07.09 16:45 A brit vállalatok mintegy
üllyedő hajón is a saját pecsenyéjükkel voltak elfoglalva. A Financial Times felmérése szerint
elyett meglévő eszközeik kihelyezésével voltak elfoglalva, ennek következtében fokozatosan és n

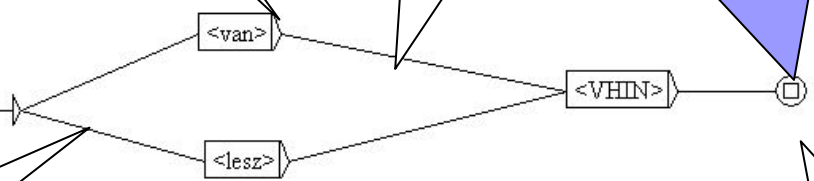
Lekérdezés gráffal

Kiinduló pont
a vizsgált szöveg tetszőleges pontja

itt vagyunk, ha a szövegben szerepel a ***van*** egy ragozott alakja

erre megyünk, ha a ***van*** után a szövegben szerepel egy **VHIN** kódú alak

A gráf csak akkor illeszkedik, azaz csak akkor van találat, ha ide el tudtunk jutni: igy vagy úgy de végig!

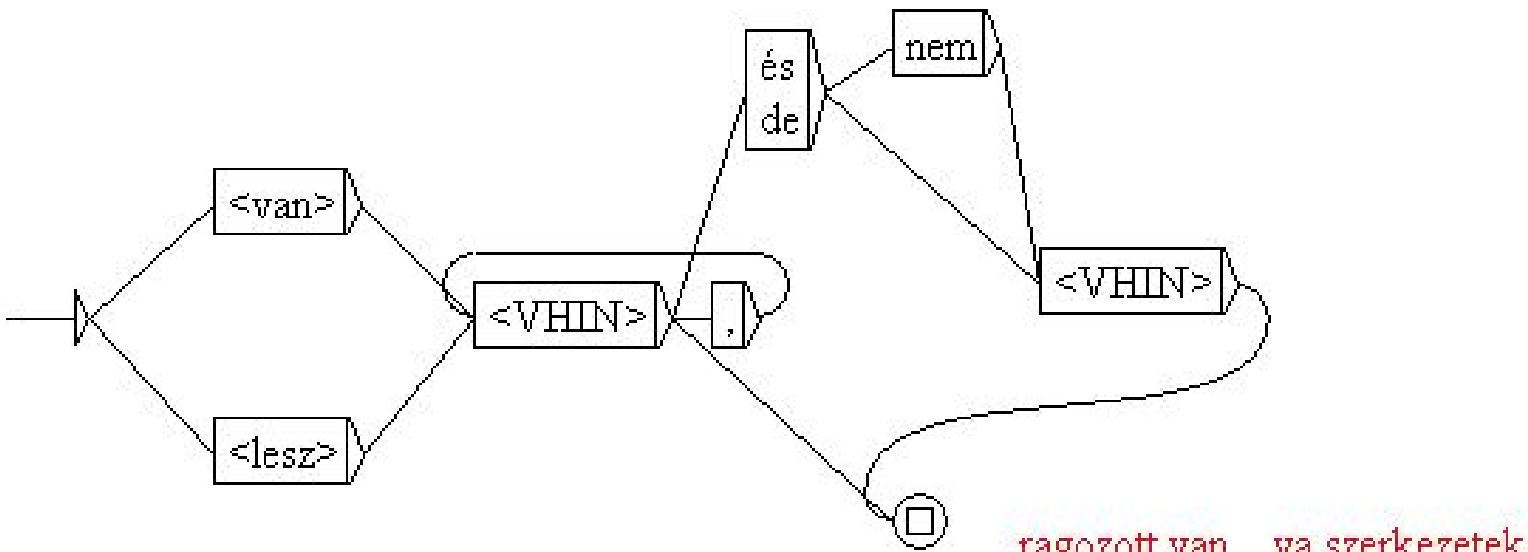


ragozott van ...va szerkezetek

a **vagy** relációt ugyannabból a pontból ugyanabba a pontba vezető több ággal jelezzük

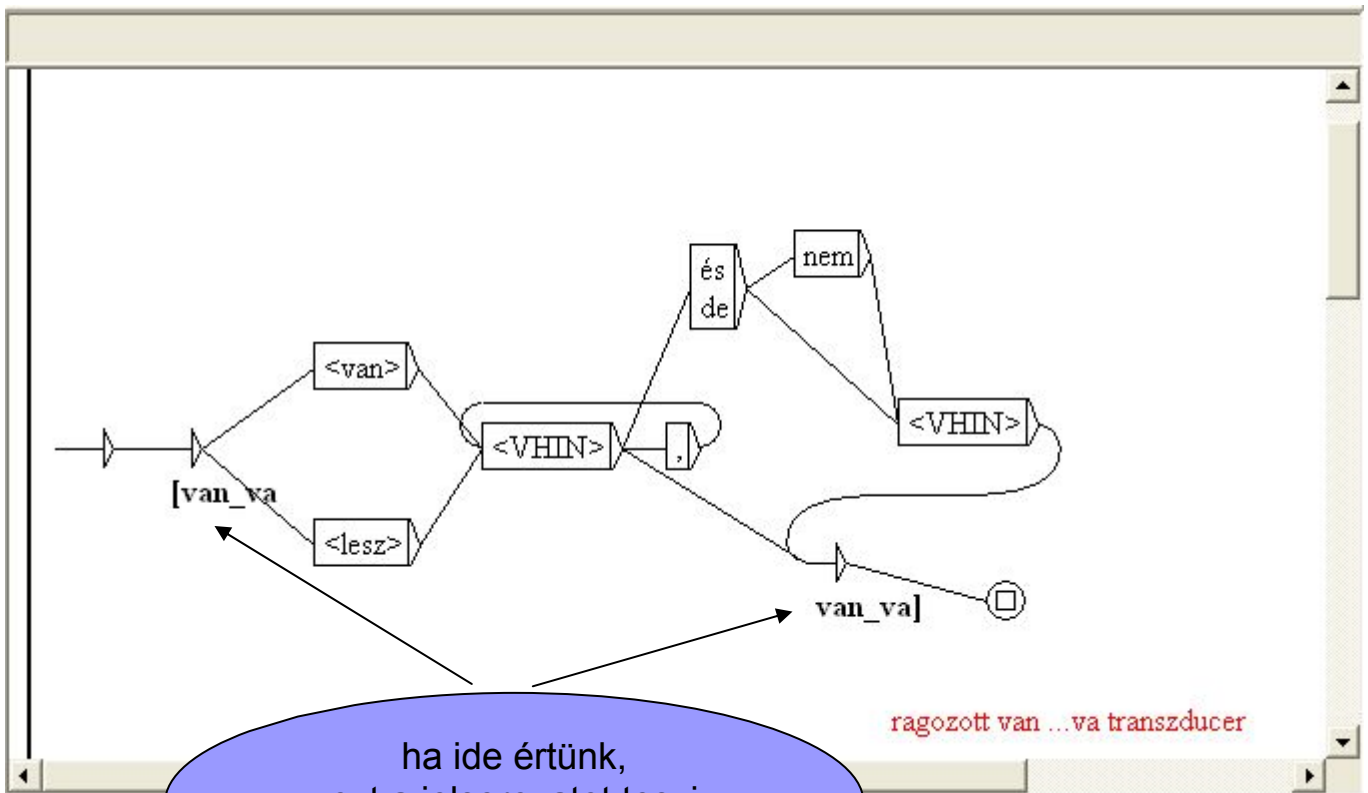
Végpont
a vizsgált szöveg azon pontja, ahova a gráf illesztése után eljutottunk

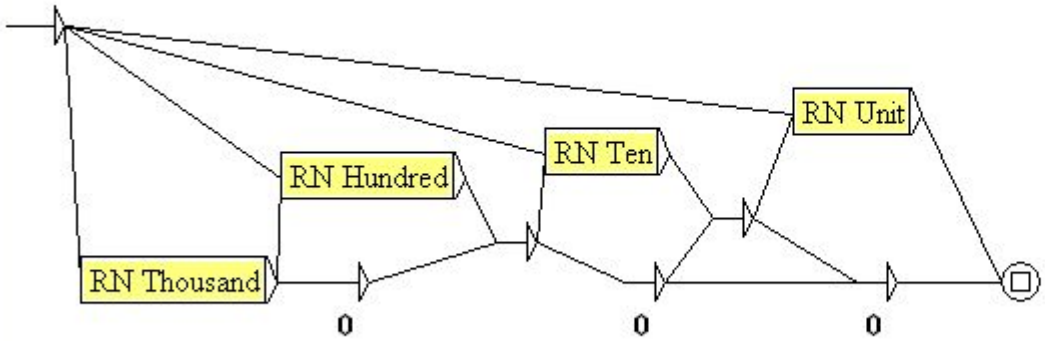
véges állapotú felismerő gráf (fsa)

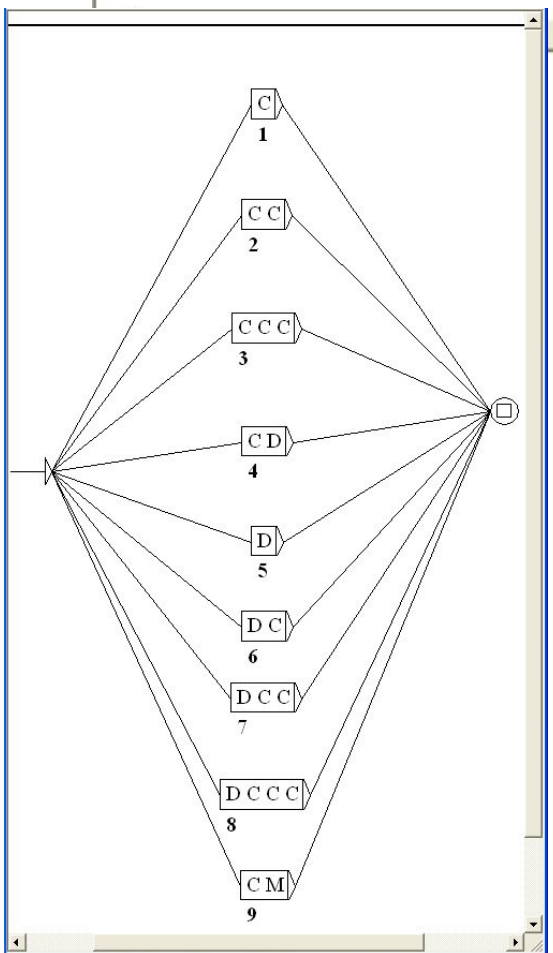


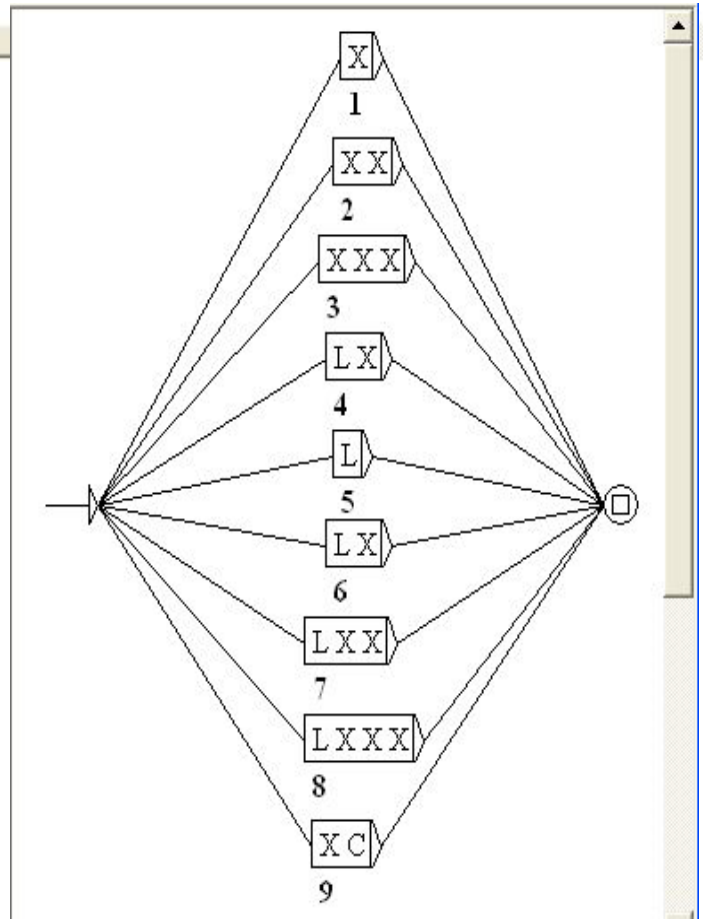
véges állapotú transzducer (fst)

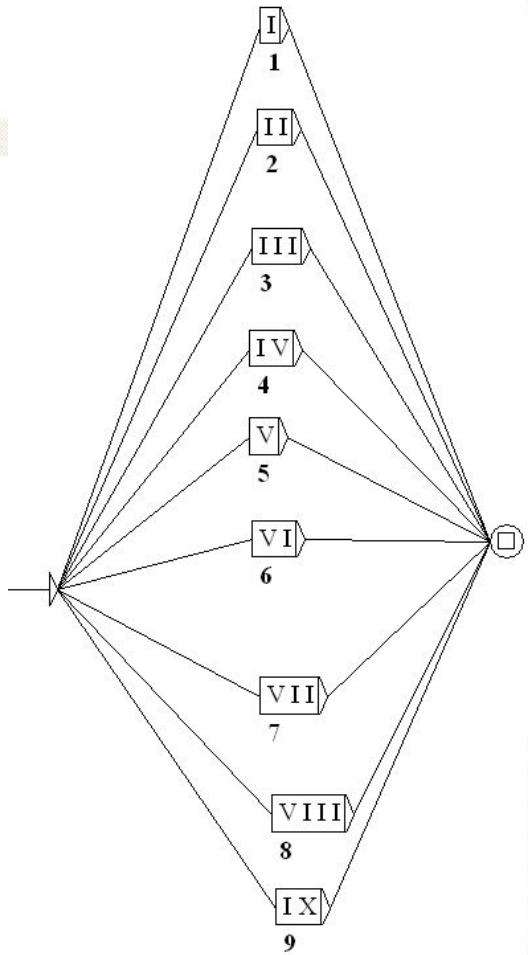
nemcsak felismer, hanem ki is ad jelsorozatot




RN Unit
RN Ten
RN Hundred
RN Thousand
0
0
0


C
1
C C
2
C C C
3
C D
4
D
5
D C
6
D C C
7
D C C C
8
C M
9

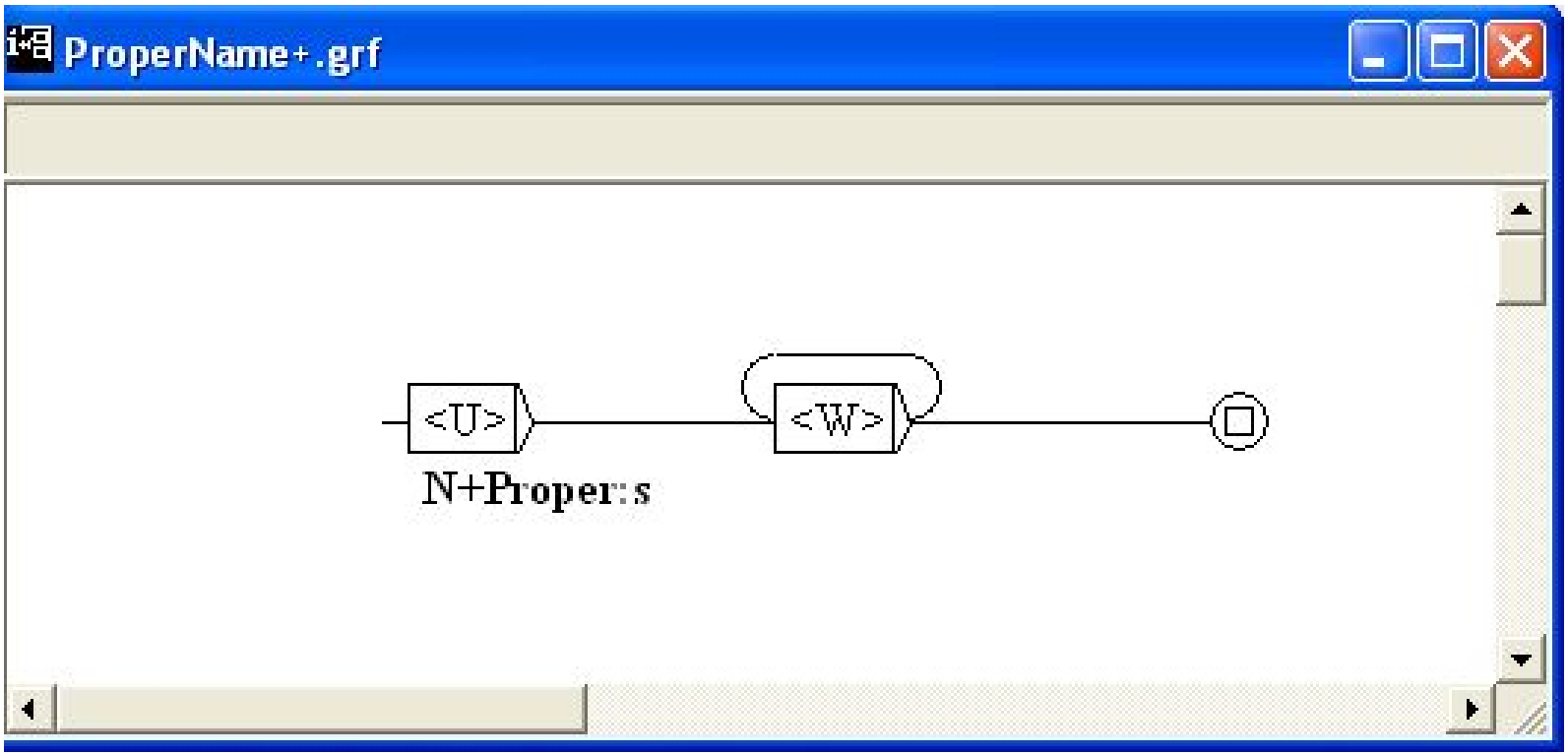

X
1
X X
2
X X X
3
L X
4
L
5
L X
6
L X X
7
L X X X
8
X C
9


I
1
I I
2
I I I
3
I V
4
V
5
V I
6
V I I
7
V I I I
8
I X
9

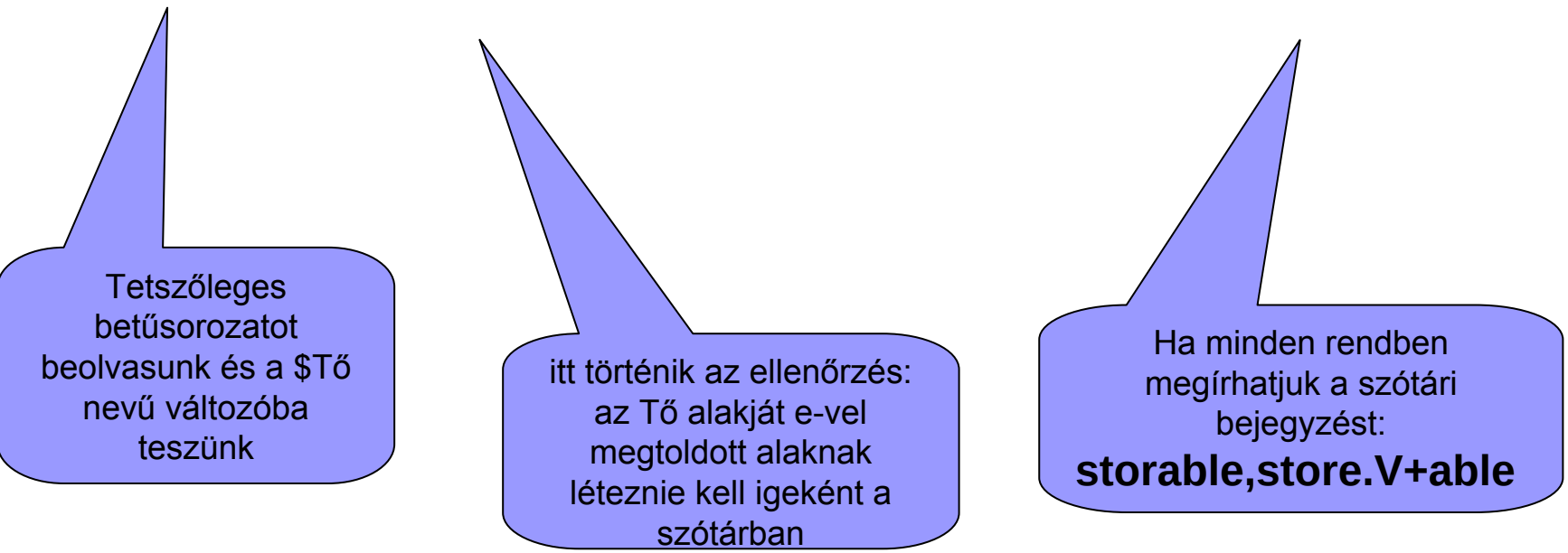
Lexikai szabályok

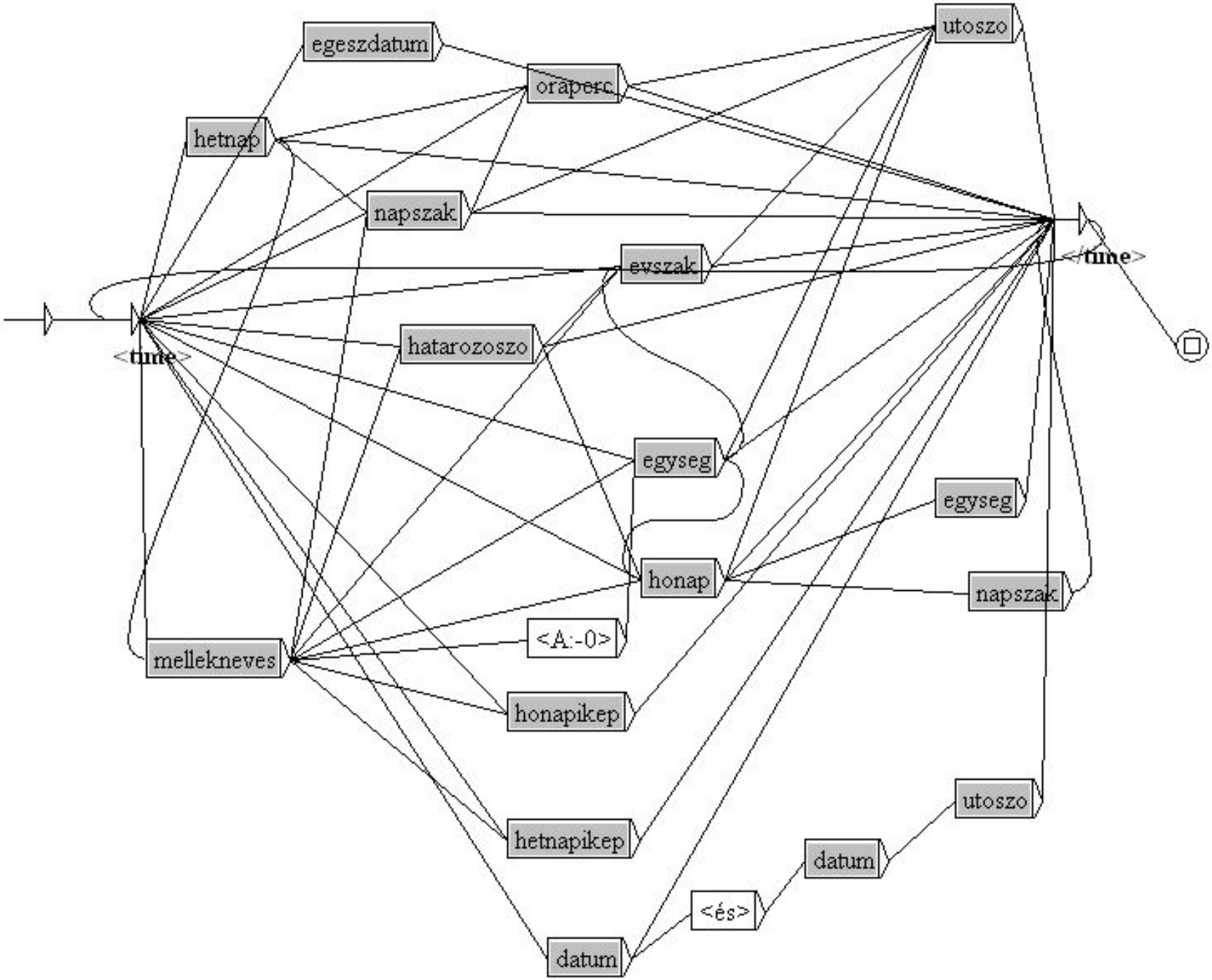
- **Lexikai transzducerek**
  - képzők, szóösszetételek elemzésére
  - tulajdonnevek felismerésére
- Lexikai megszorításokkal ellenőrizhetjük az elemek érvényességét
- Kimenetük új szótári elemek, melyeket hozzátehetünk az alapszótárhoz

végső tulajdonnév szabály



lexikai megszorítások




egeszdatum
oraperc
utoszo
hetnap
napszak
<time>
evszak
<time>
hatarozoszo
egyseg
egyseg
honap
napszak
<A:-0>
mellekneves
honapikep
hetnapikep
utoszo
datum
<és>
datum

Magyar változat

- Kulcskérdés: a lexikai modul elkészítése
- szótár egyben morfológiai elemzést is ad
- a magyar morfológia listázással nem megoldható
- Kompromisszum:
  óriási korpuszok elemzett szókincse
  HUMOR morfológiai elemzővel elemezve
  - Magyar Nemzeti Szövegtár 150 m szó
  - Magyar webkorpusz (MOKK) 2500m szó

Az INTEX korlátai

- Gyors, hatékony, de sok beépített korlát
  - egyszerre egy nyelvvel lehet dolgozni
  - zárt rendszer – könnyű bevinni szövegeket, de nehéz kivinni őket
  - nincs xml kimenet
  - egyszerre egyetlen szöveggel lehet dolgozni
  - a szöveg annotációja lépcsőzetes elemzésnél nehézkes

JÖN! NooJ!

- Teljesen átdolgozott, többnyelvű, rugalmas, áttekinthetőbb rendszer
- INTEX utódja – elveiben azonos, de továbbfejlesztett változat
- Első nyilvános változat: december 15-én!
- A magyar változat hónapok óta előkészületben

Magyar NooJ

- Működő morfológiai elemző több változatban
  - nyers erő
  - morfológiai gráfok jegyek alkalmazásával
  - online elemző szótár paradigmatáblák használatával
- Az Értelmező Kéziszótár szókincse teljeskörű ragozással

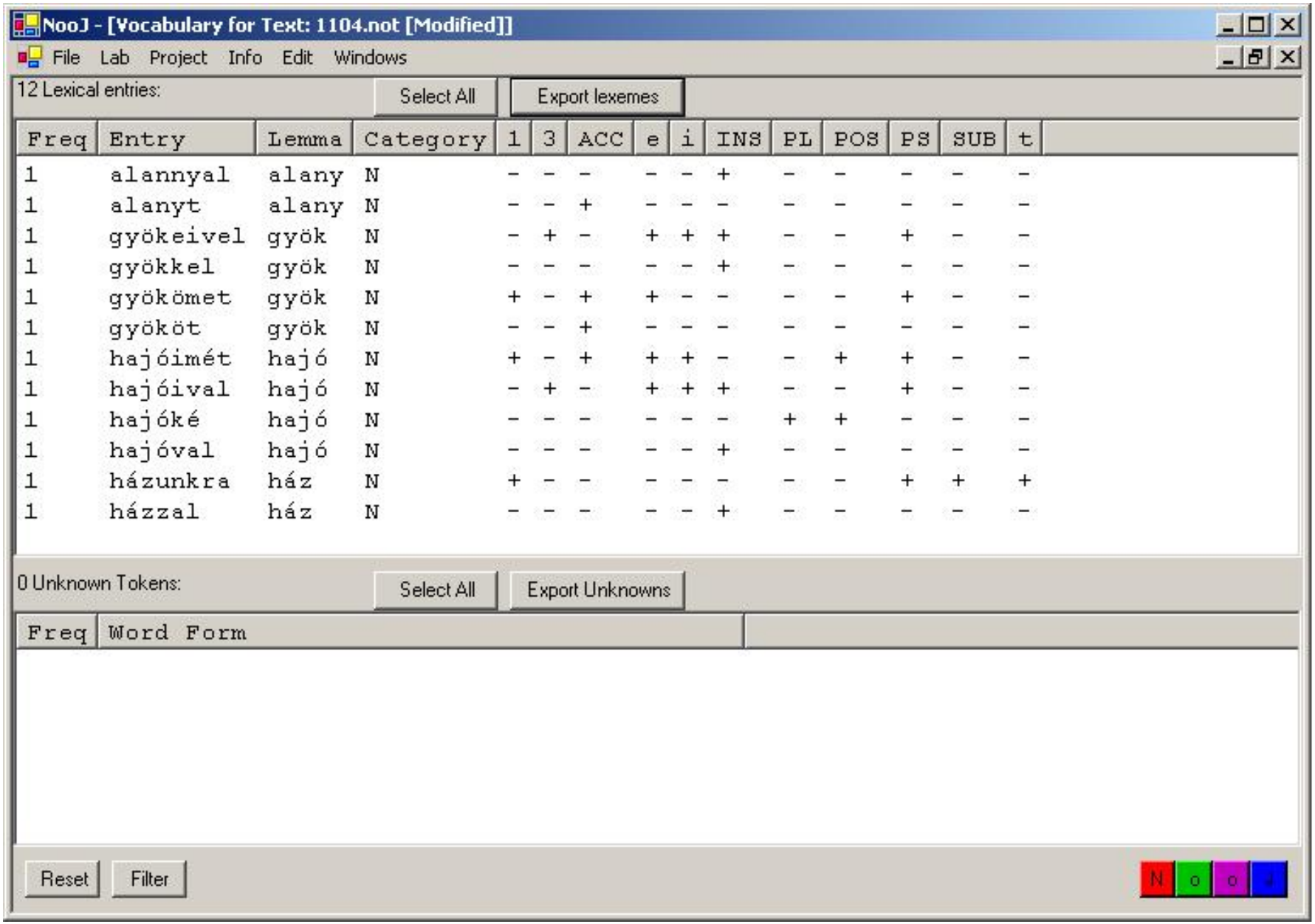
NooJ - [news.dic]

File Lab Project Info Edit Windows

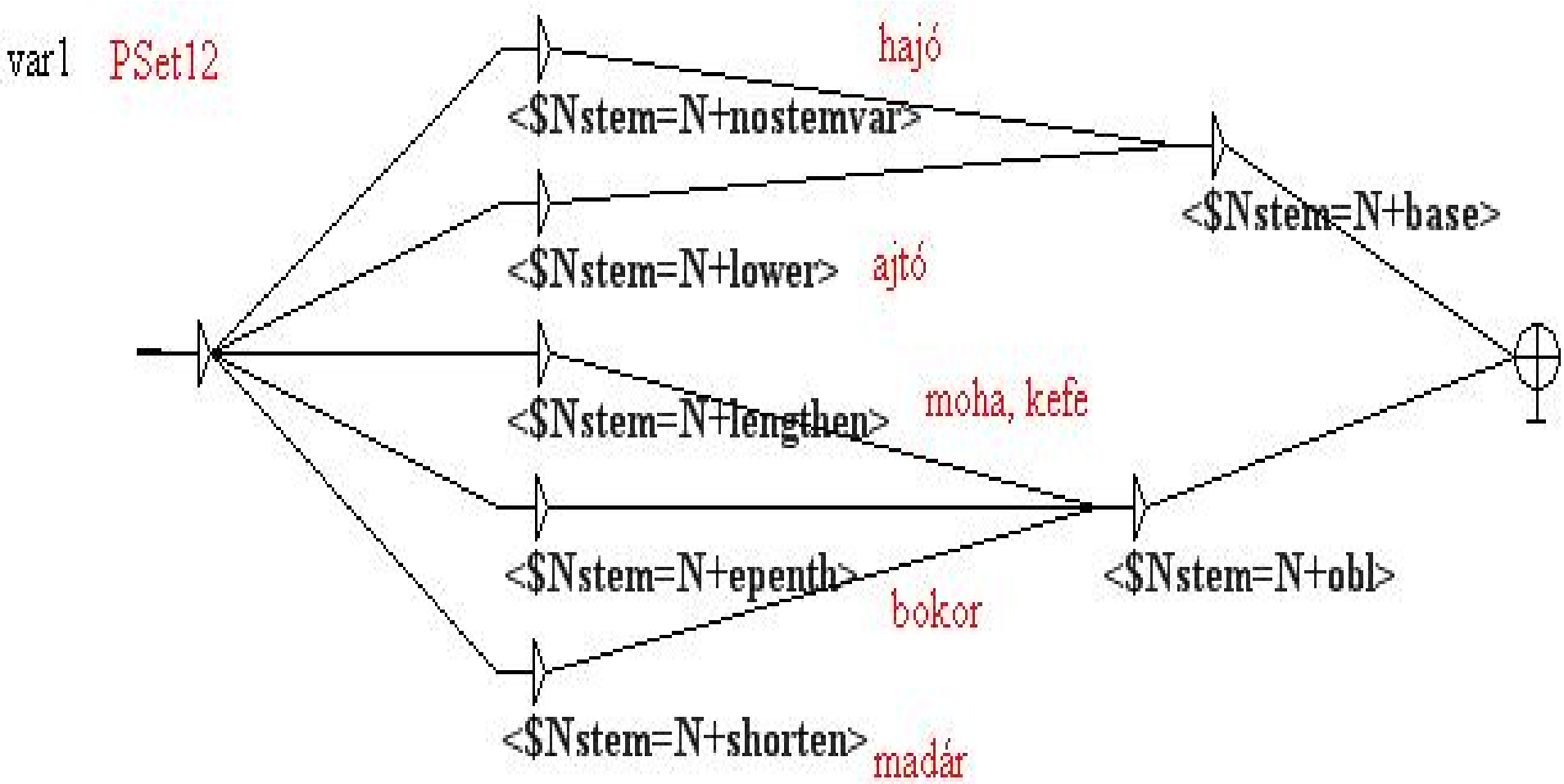
Dictionary contains 59213 entries

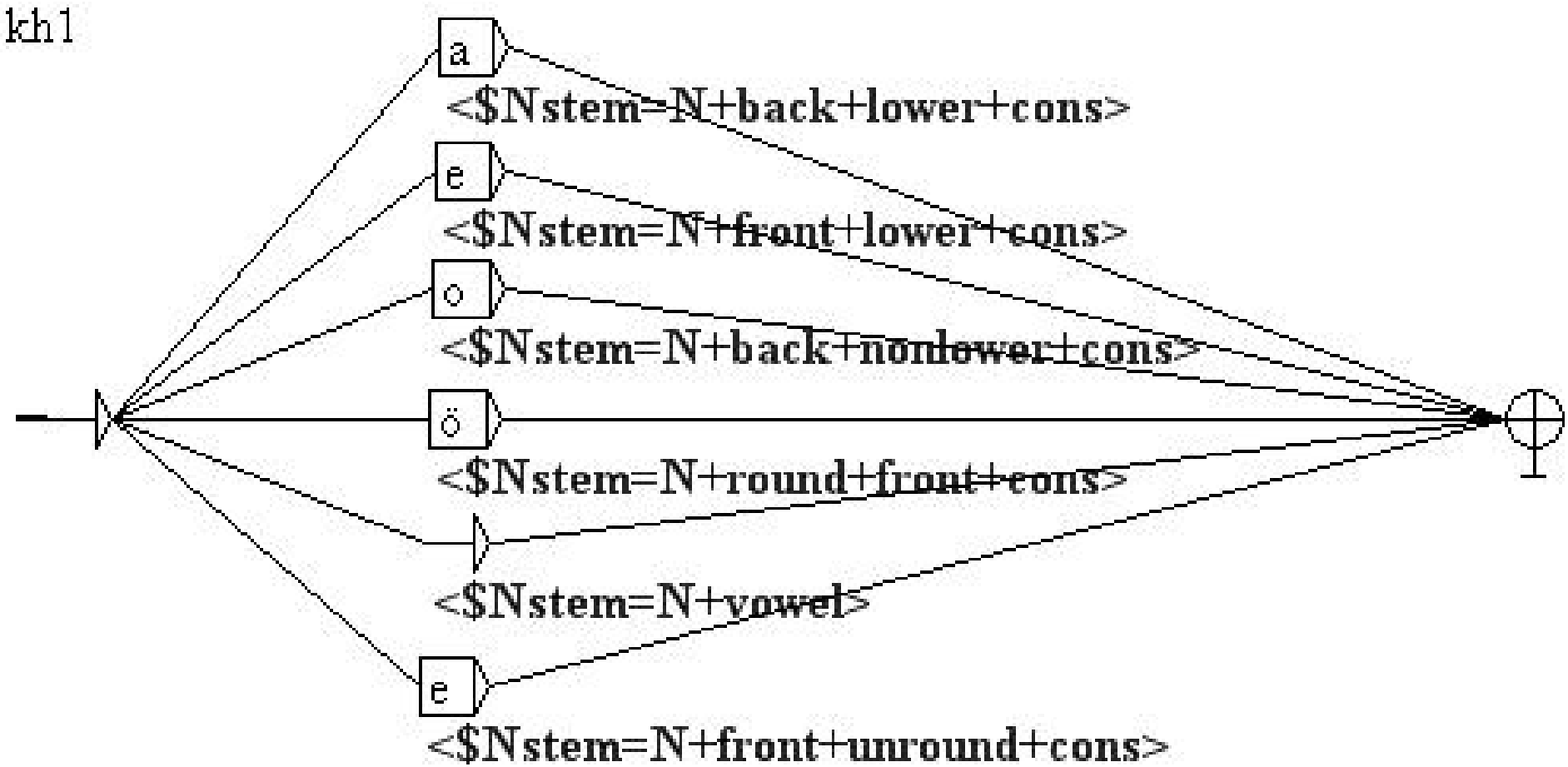
```
# NooJ V1
# Dictionary
#
# Input Language is: hu
#
# Alphabetical order is not required.
# Number of fields must be constant.
#
# Use any number of inflectional description files (.nof or .nog)
# Special Command: #use description.nof
# Special Command: #use description.nog
# DESCRIPTION FILES MUST BE STORED IN SAME DIRECTORY AS DICTIONARY
#
# Special Features: +FLX +UNAMB
#
# Special Characters: '\' '"' '+' ',' '#' ' '
#

abaposztó,N+base+nostemvar+vowel+back+j+accnolink+nonlower
abba,PRO+base+nostemvar+vowel+back+j+accnolink+nonlower
abban,PRO+base+nostemvar
abbé,N+base+nostemvar+vowel+back+j+accnolink+nonlower
abbeli,A+base+nostemvar+vowel+front+unround+j+accnolink+nonlower
abból,PRO+base+nostemvar
abbreviatúrá,N+obl+shorten+vowel+back+j+accnolink+nonlower
abbreviatúra,N+base+shorten+vowel+back+j+accnolink+nonlower
ábécé,N+base+nostemvar+vowel+front+unround+j+accnolink+nonlower
ábécérend,N+base+nostemvar+cons+front+unround+j+acclink+nonlower
ábécés,A+base+nostemvar+cons+front+unround+nonj+accnolink+nonlower
ábécés,N+base+nostemvar+cons+front+unround+nonj+accnolink+nonlower
ábécéskönyv,N+base+nostemvar+cons+front+round+nonj+acclink+nonlower
aberráció,N+base+nostemvar+vowel+back+j+accnolink+nonlower
ablak,N+base+nostemvar+cons+back+j+acclink+nonlower
ablakbélés,N+base+nostemvar+cons+front+unround+nonj+accnolink+nonlower
ablakbiztosítás,N+base+nostemvar+cons+back+nonj+accnolink+nonlower
ablakdeszká,N+obl+shorten+vowel+back+j+accnolink+nonlower
ablakdeszka,N+base+shorten+vowel+back+j+accnolink+nonlower
```


NooJ - [Vocabulary for Text: 1104.not [Modified]]
File Lab Project Info Edit Windows
12 Lexical entries:
Select All
Export lexemes
Freq Entry Lemma Category 1 3 ACC e i INS PL POS PS SUB t
1 alannyal alany N - - - - - + - - - - -
1 alanyt alany N - - + - - - - - - - -
1 gyökeivel gyök N - + - + + + - - + - -
1 gyökkel gyök N - - - - - + - - - - -
1 gyökömet gyök N + - + + - - - - + - -
1 gyököt gyök N - - + - - - - - - - -
1 hajóimét hajó N + - + + + - - + + - -
1 hajóival hajó N - + - + + + - - + - -
1 hajóké hajó N - - - - - - + + - - -
1 hajóval hajó N - - - - - + - - - - -
1 házunkra ház N + - - - - - - - + + +
1 házzal ház N - - - - - + - - - - -
0 Unknown Tokens:
Select All
Export Unknowns
Freq Word Form
Reset
Filter
N o o

A tővariánsjegyek ellenőrzése




kh1
a
<$Nstem=N+back+lower+cons>
e
<$Nstem=N+front+lower+cons>
o
<$Nstem=N+back+nonlower+cons>
ö
<$Nstem=N+round+front+cons>
<$Nstem=N+vowel>
e
<$Nstem=N+front+unround+cons>

Összegzés

- INTEX/NooJ általános célú integrált nyelvelemző rendszer
- Nemcsak számítógépes nyelvészeknek …
- http://corpus.nytud.hu/INTEX
- Közreadjuk a magyar nyelvi eszközeinket a szakmai nagyközönség számára
- Várunk érdeklődőket, önkéntes segítőket, a kutatási eredmények megosztását

Végül, de nem utolsósorban…

- Külön köszönet kollégáimnak, akiknek a munkájáról szólt ez a beszámoló:

  - Gábor Kata
  - Nagy Viktor
  - Vajda Péter
  - Sass Bálint
  - Oravecz Csaba
  - Dancsecs Erzsébet
  - Mészáros Ágnes
  - Héja Enikő