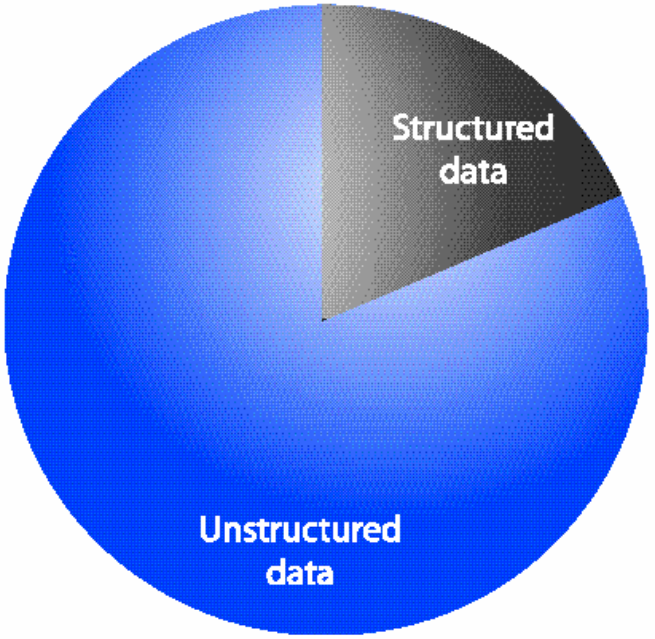
Magyar nyelvű szöveganalitika

Körmendi György
SPSS Hungary

2007 november 6.

Mit várunk a szöveganalitikától?

- A vállalatoknál tárolt információk 80%-a struktúrálatlan adatokban /IDC/
- Információ kinyerése!
- Hol tart?

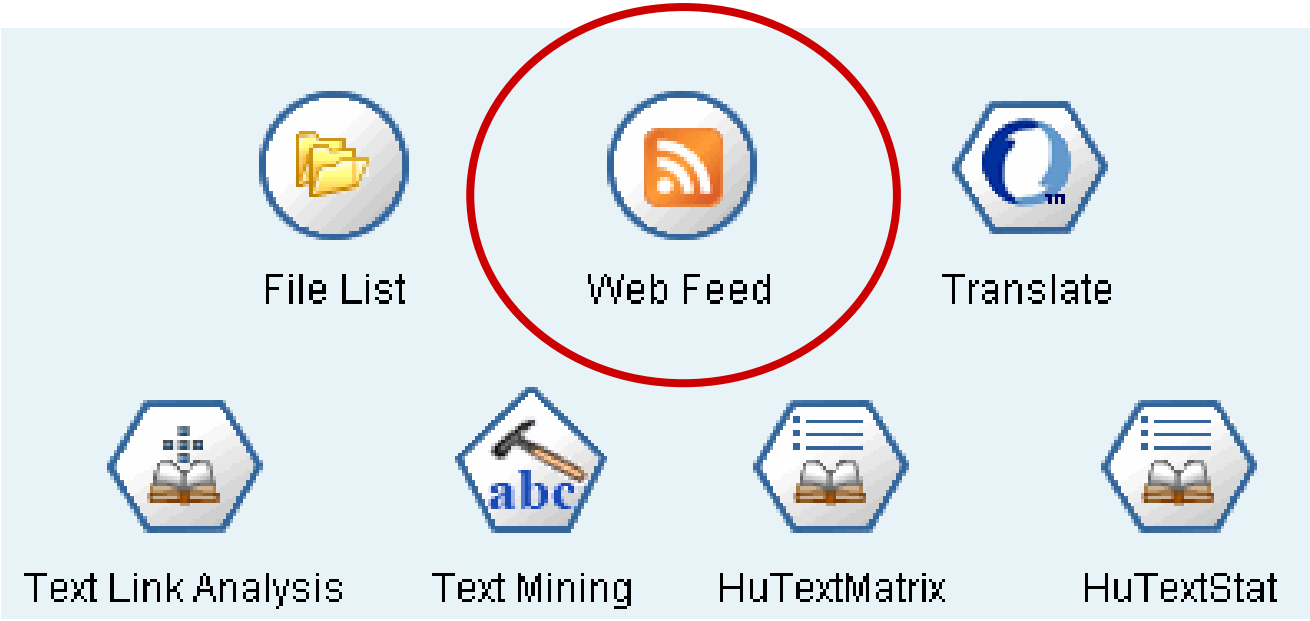


Mit NEM tárgyalunk?

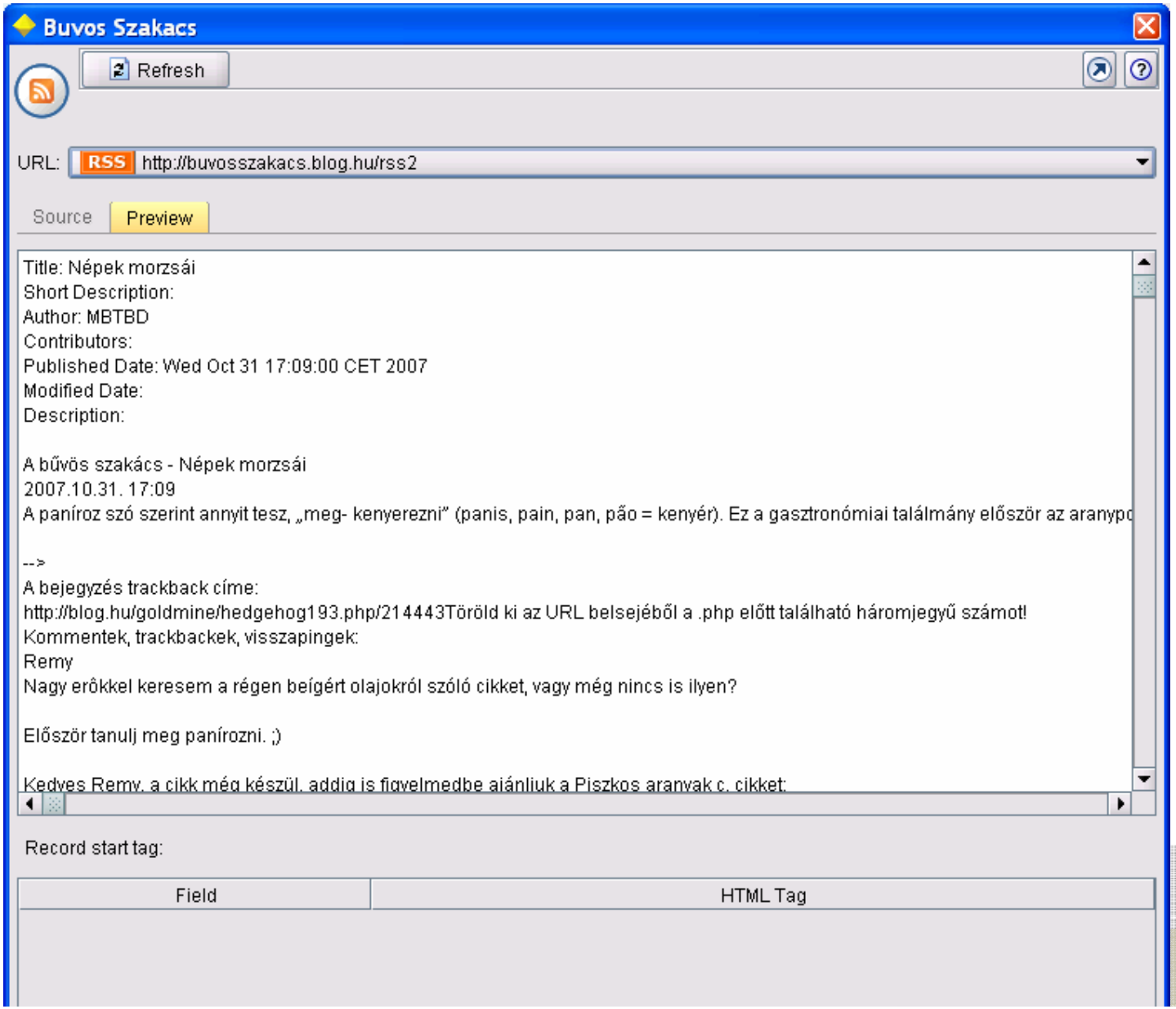
- Keresés
- Speciális feladatok (pl. spamszűrés)

Text analitikai eszköztár TMFC 5.0

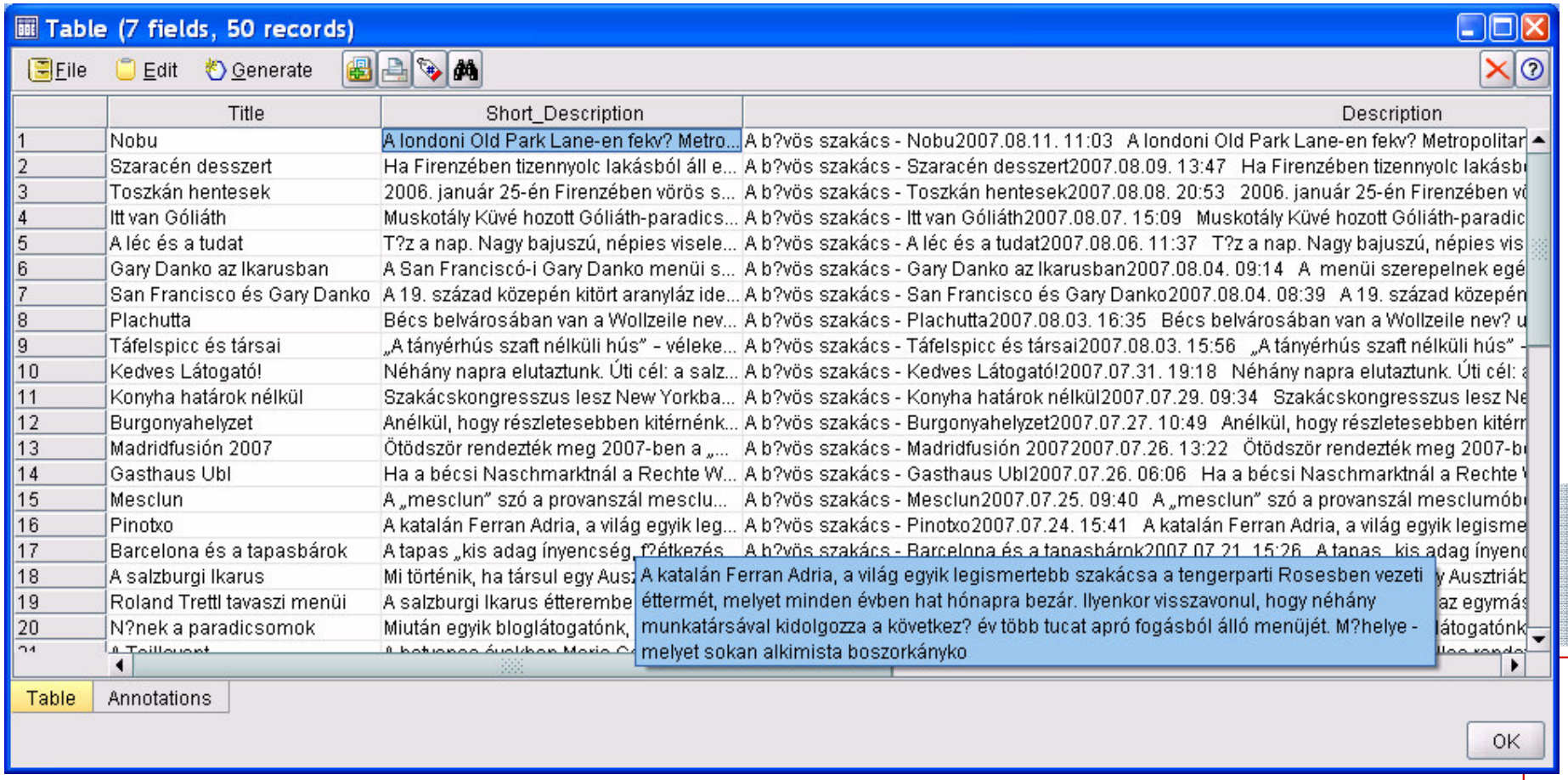
Web feed node



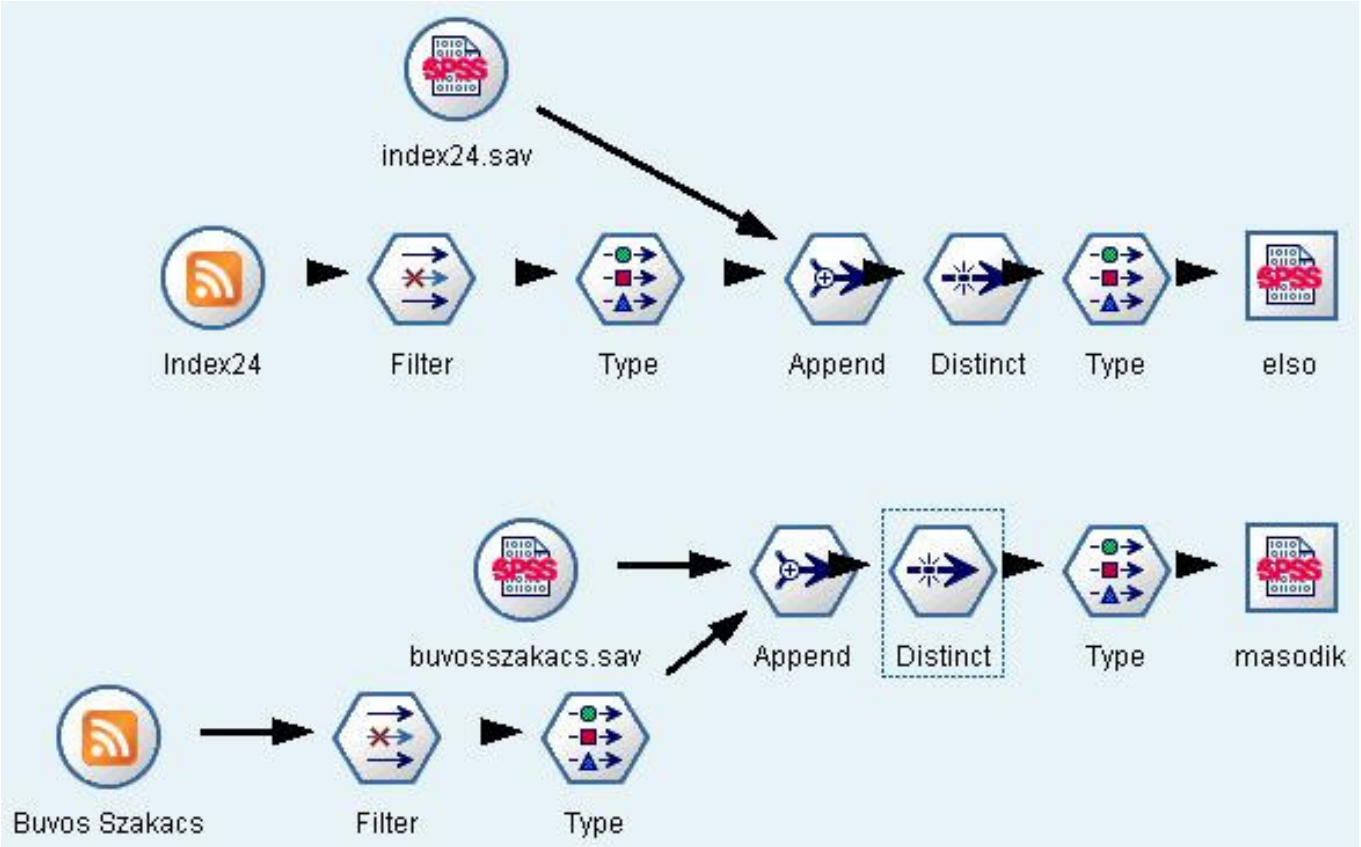
Web feed node



Web feed node



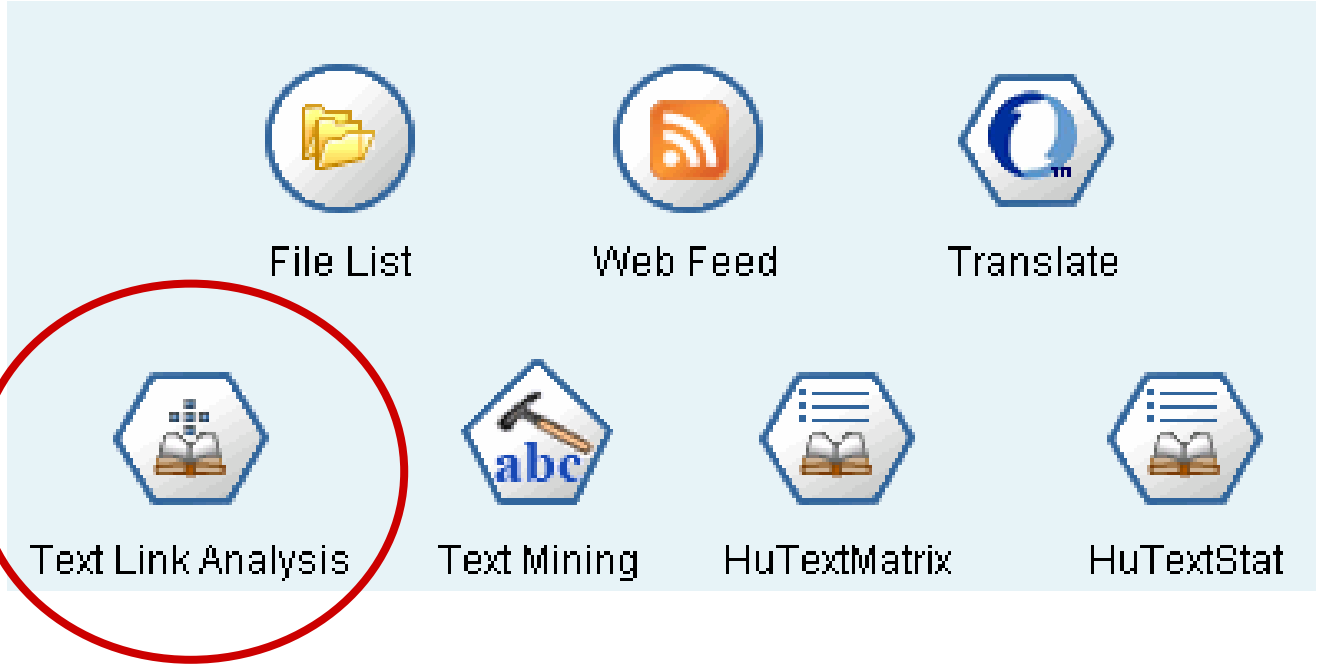
Web feed node



Statisztikai gépi fordító – Language Weaver

A textanalitika központi eleme

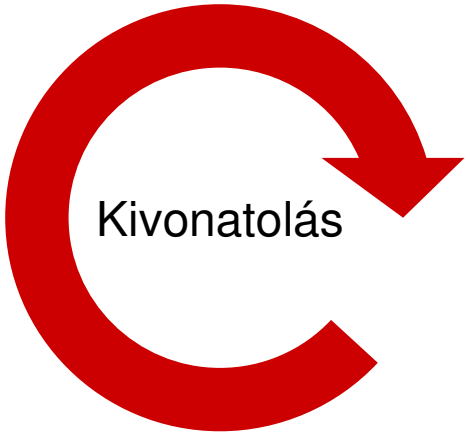
Különálló Text Link Analysis



Magyar nyelvű kivonatolás

Szemantikus szöveganalitika

Szótárak, Könyvtár

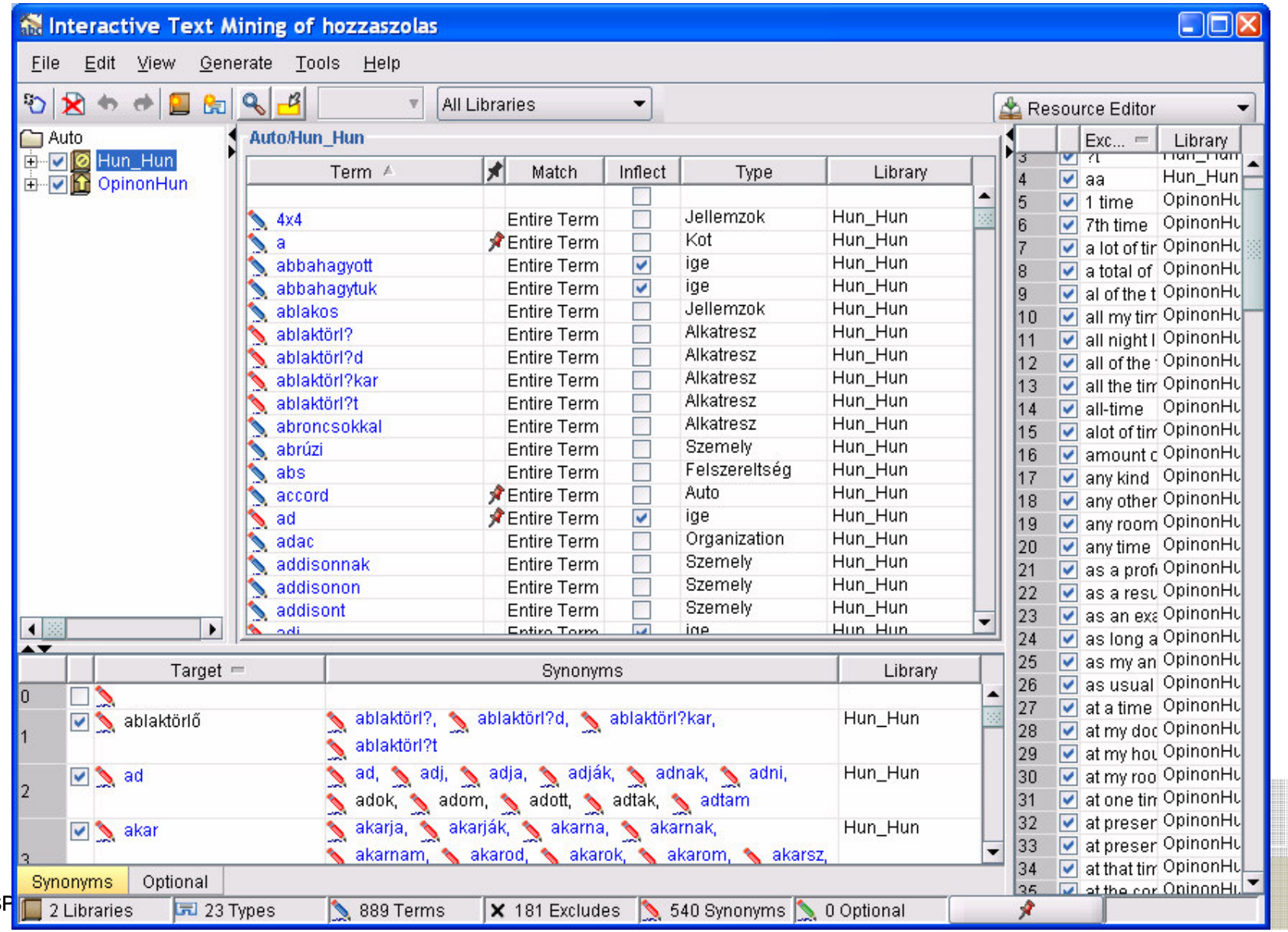


Az ipafai papnak fapipája van.
A Clementine a legjobb adatbányász szoftver.

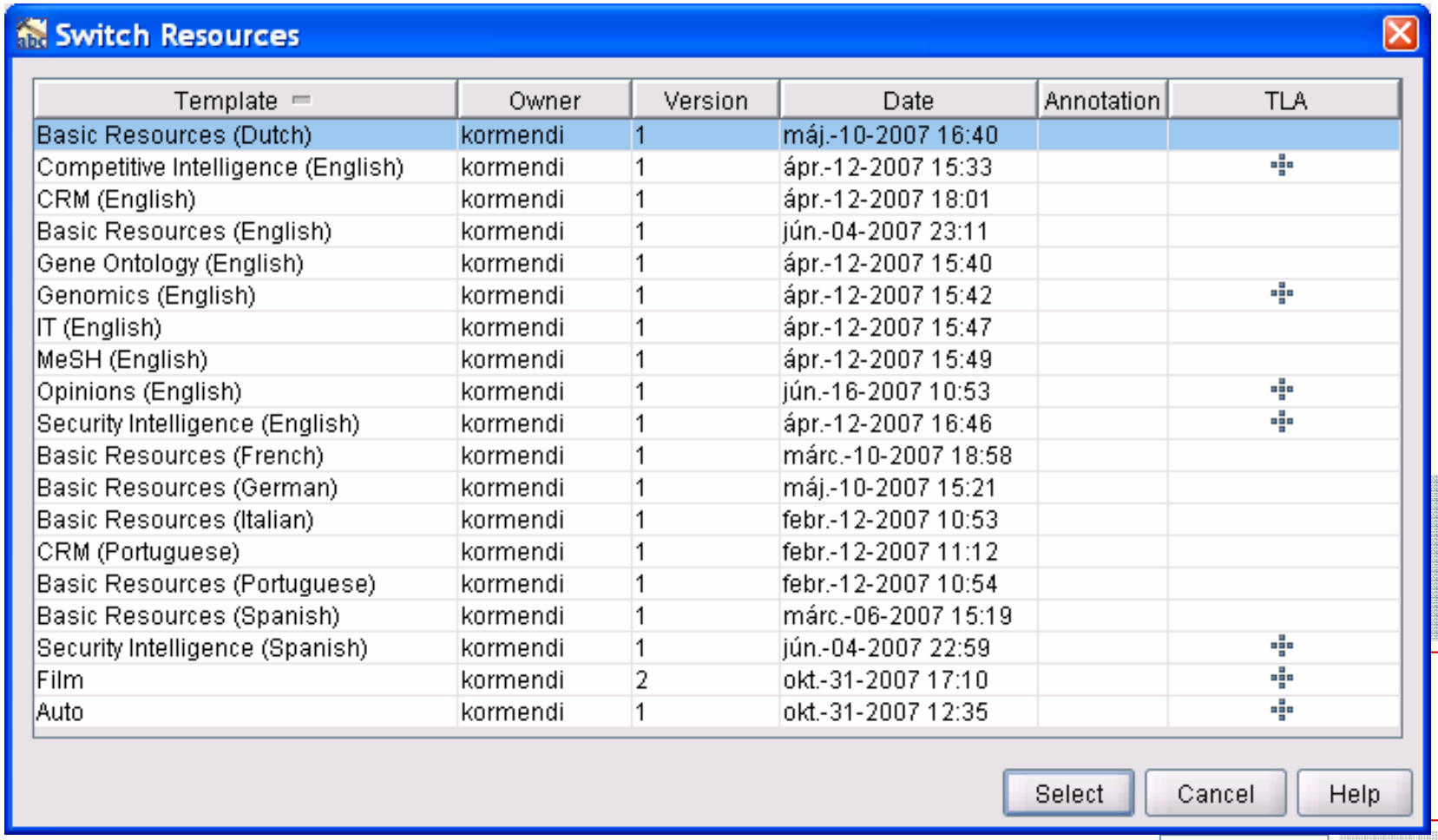
Nyelvi minták azonosítása

A textanalitika központi eleme

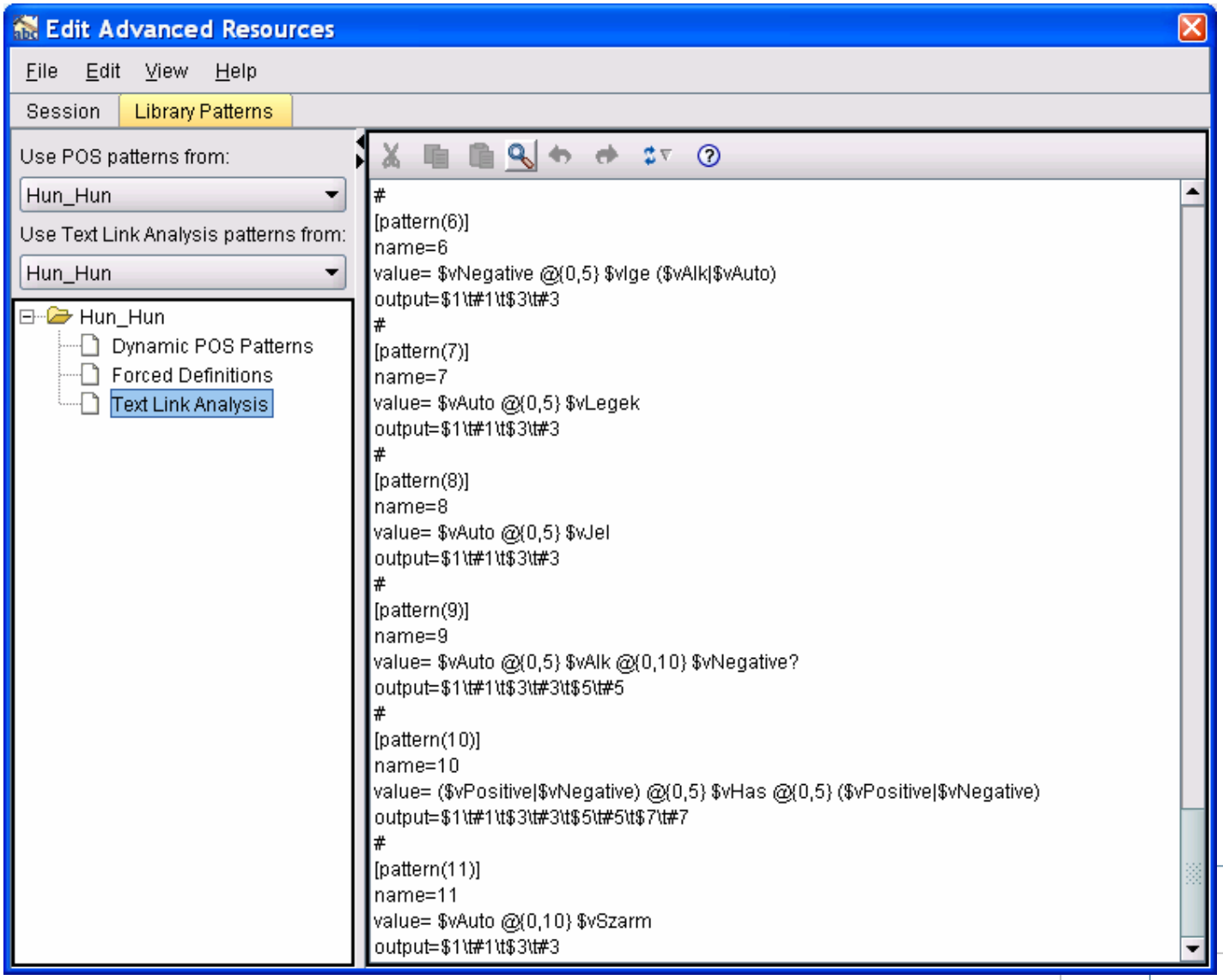
Szótár



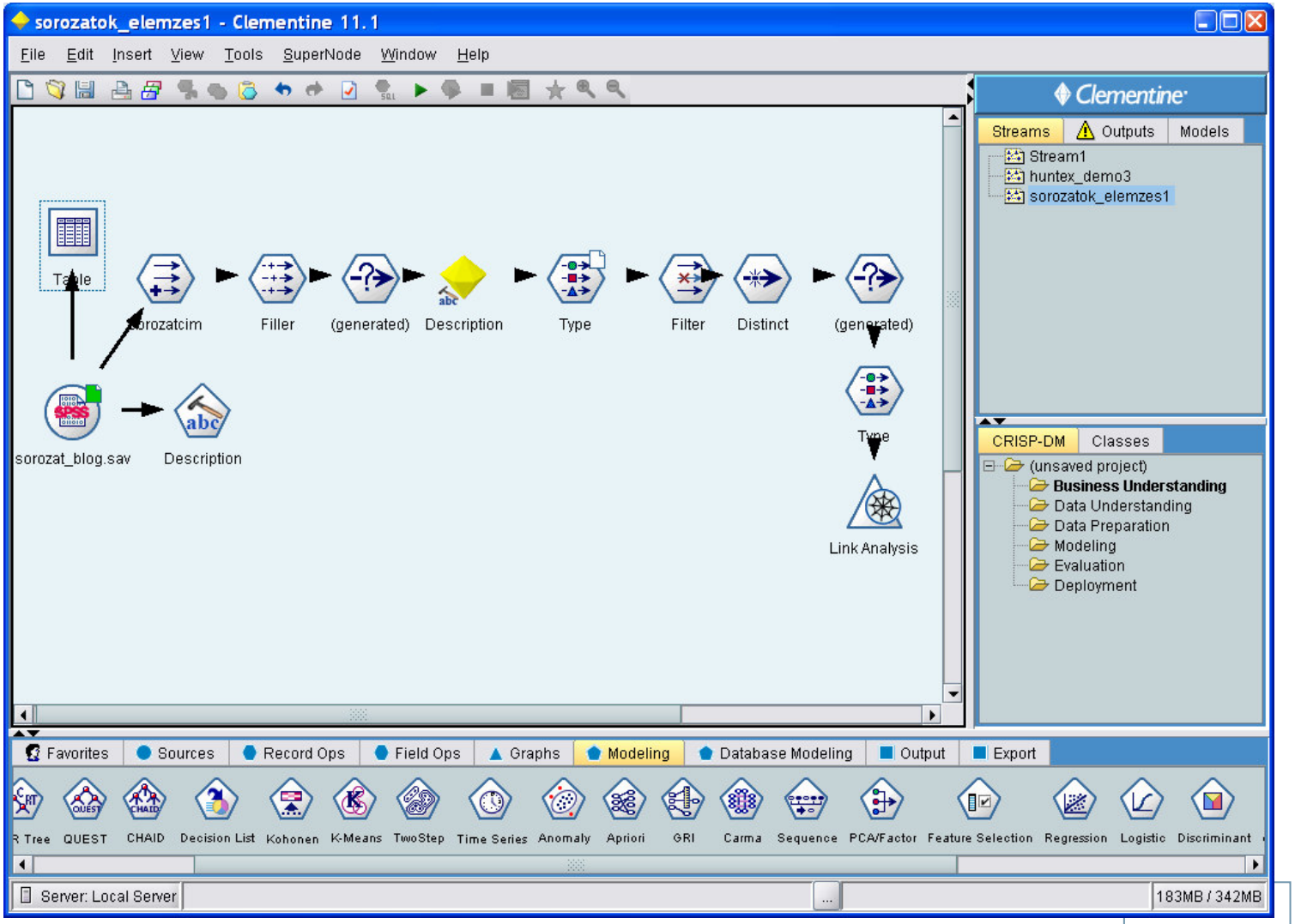
Szótár



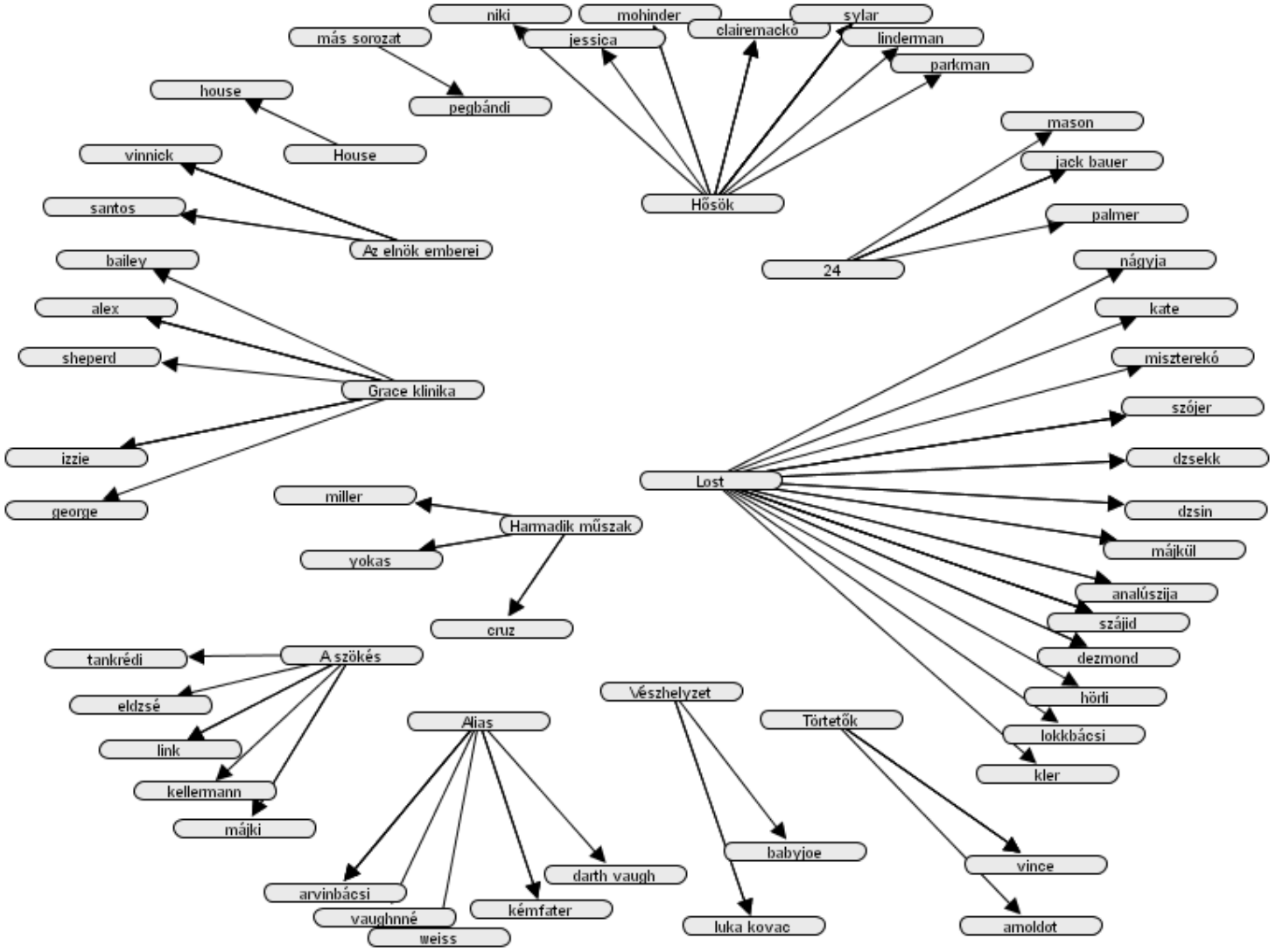
Text Link Analysis



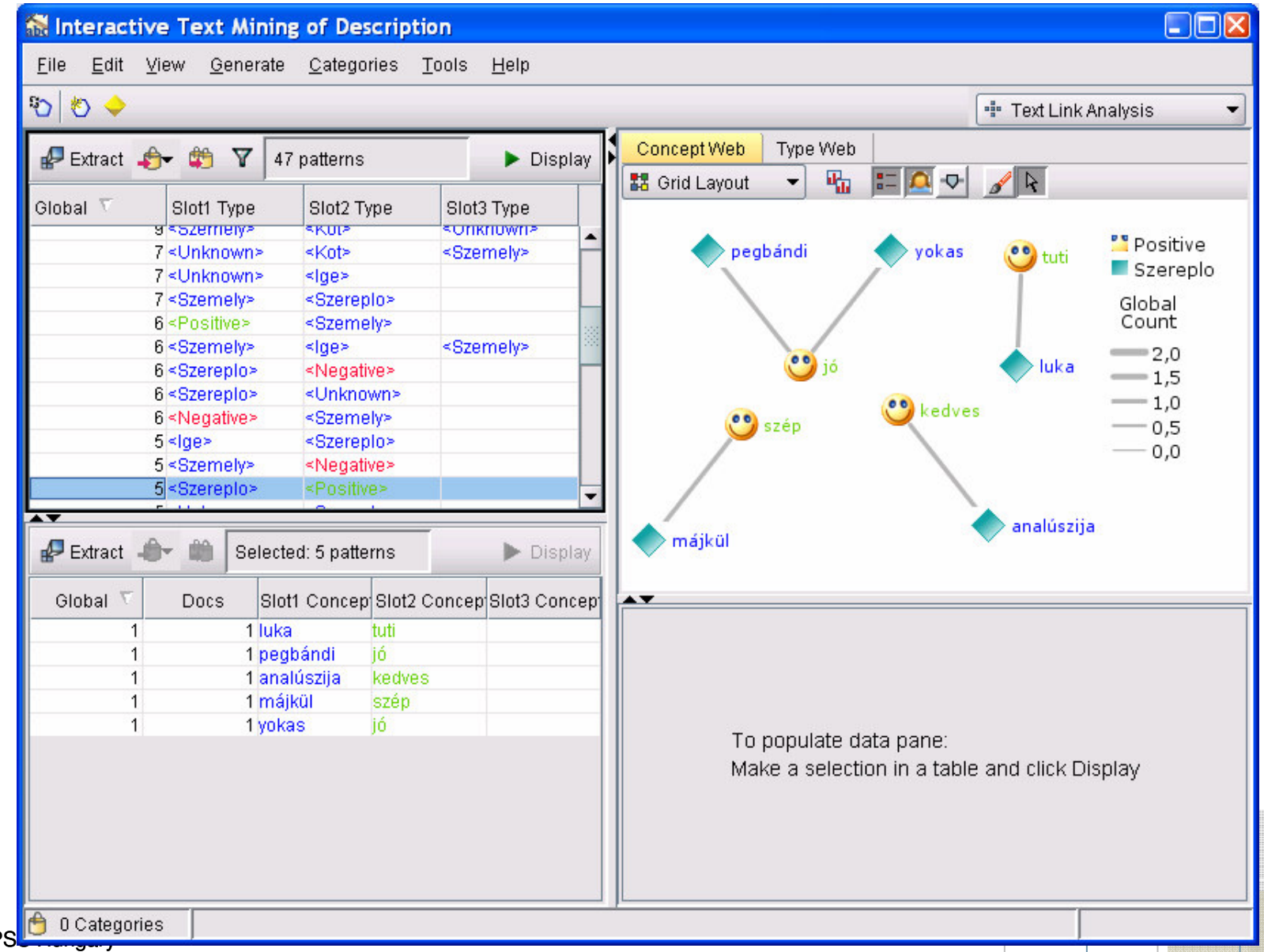
Magyar példák



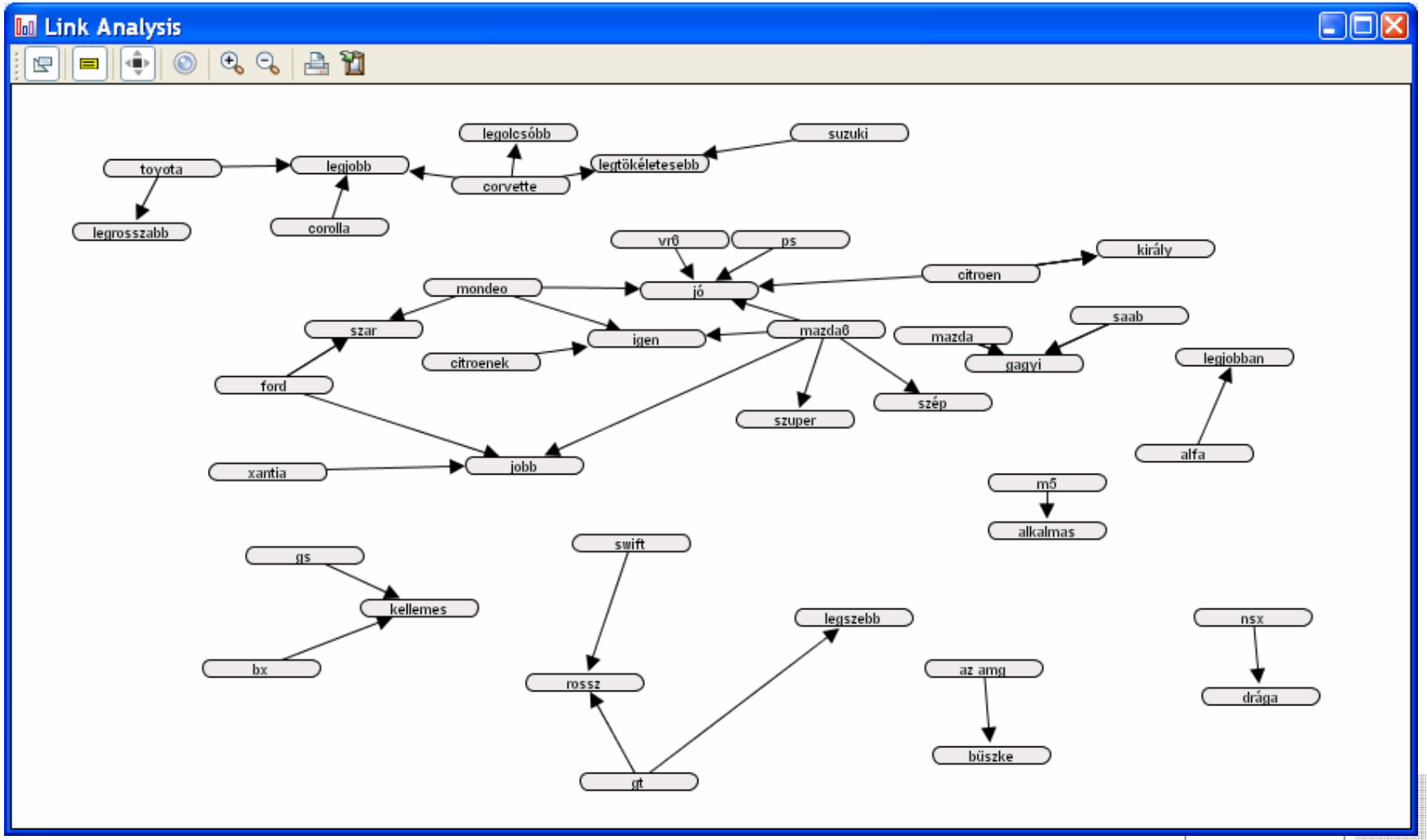
Magyar példák



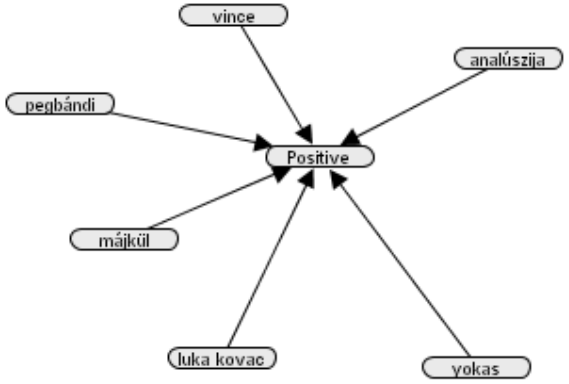
Text link analysis

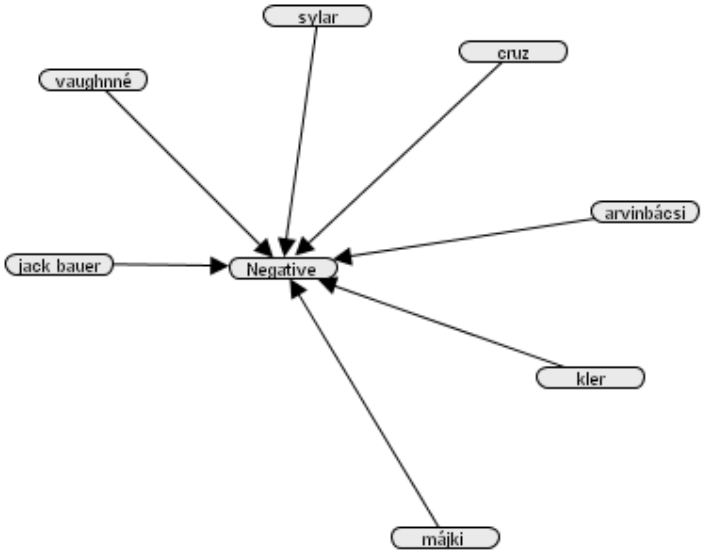


Magyar példák

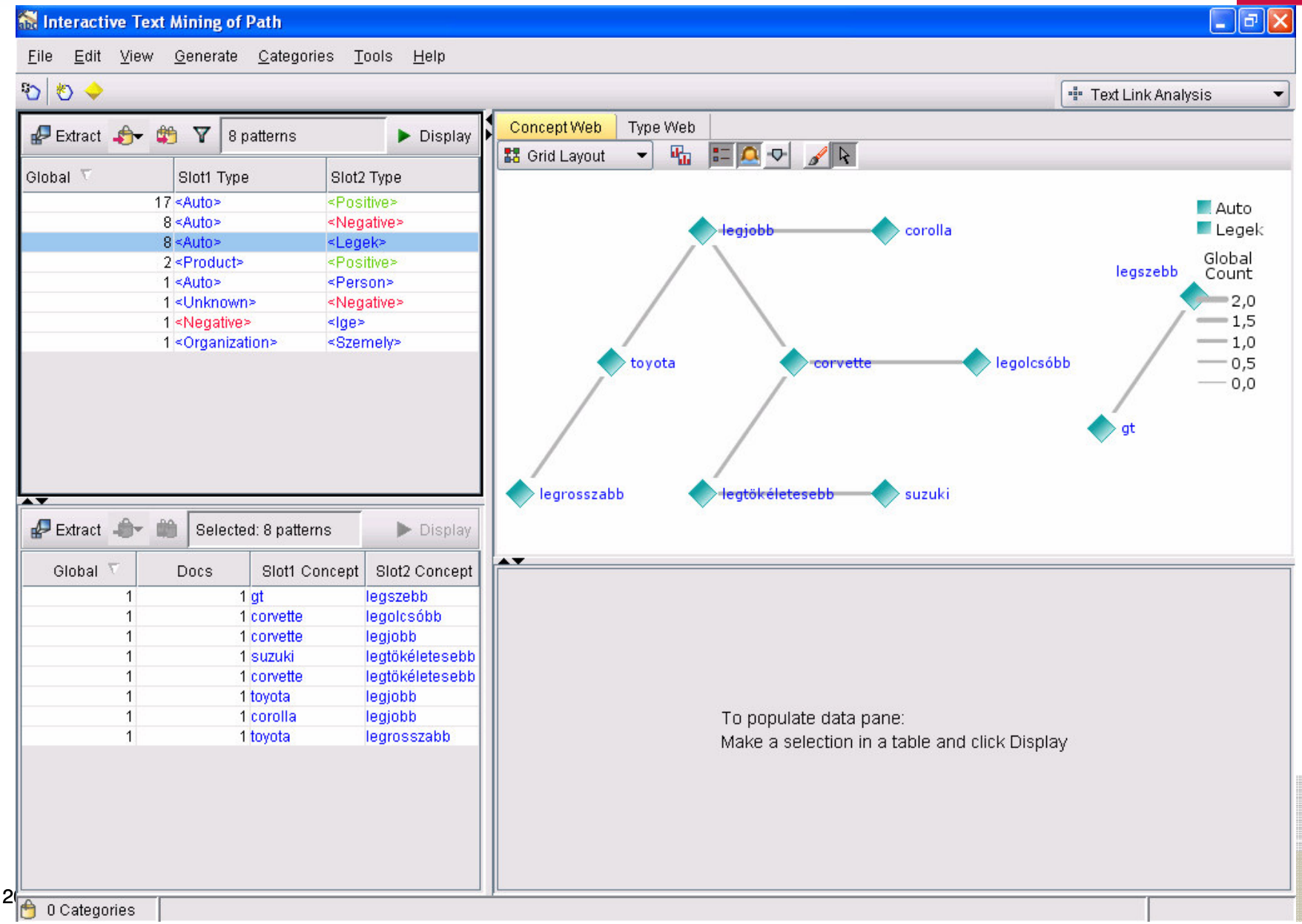


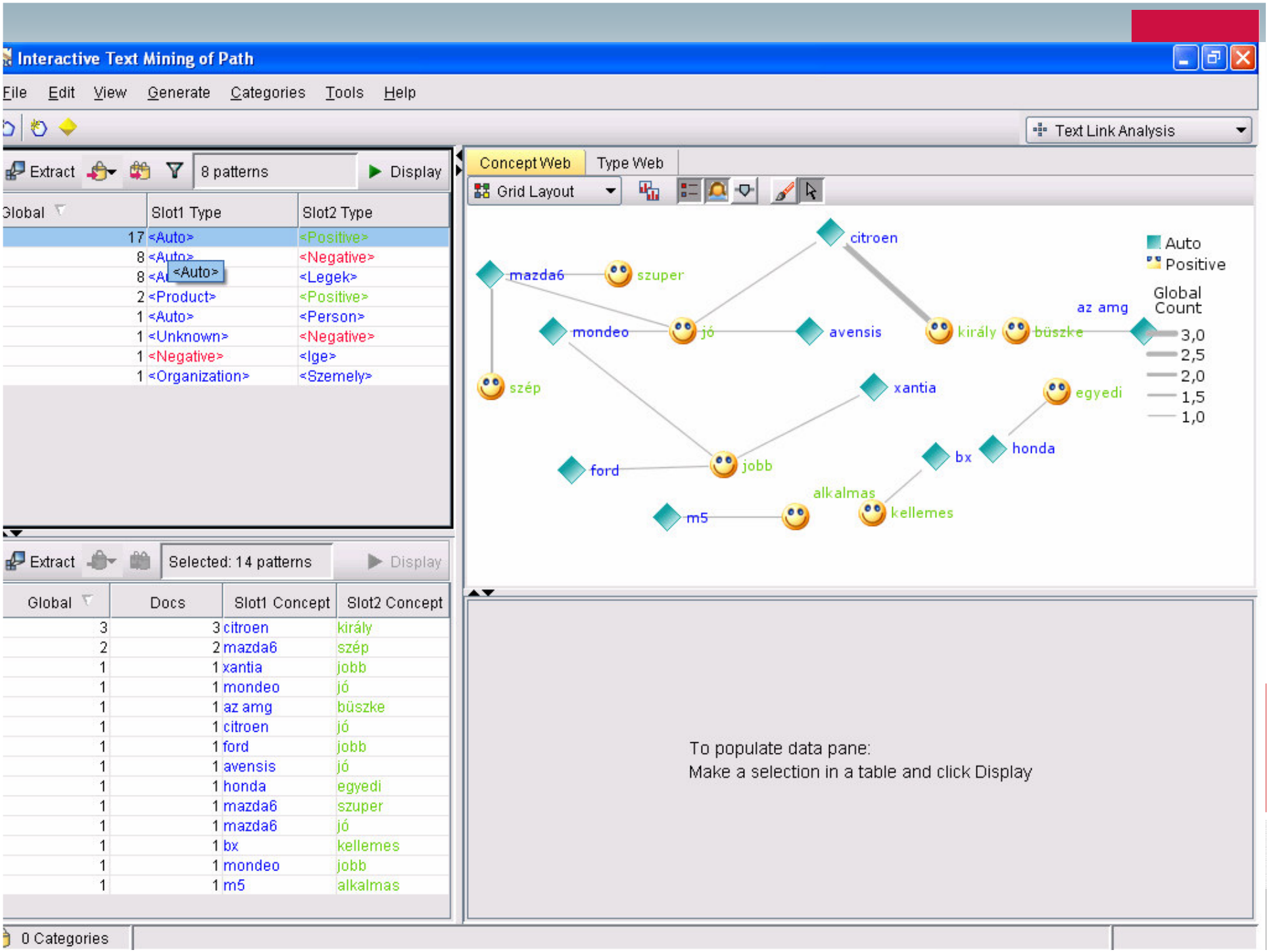
Magyar példák

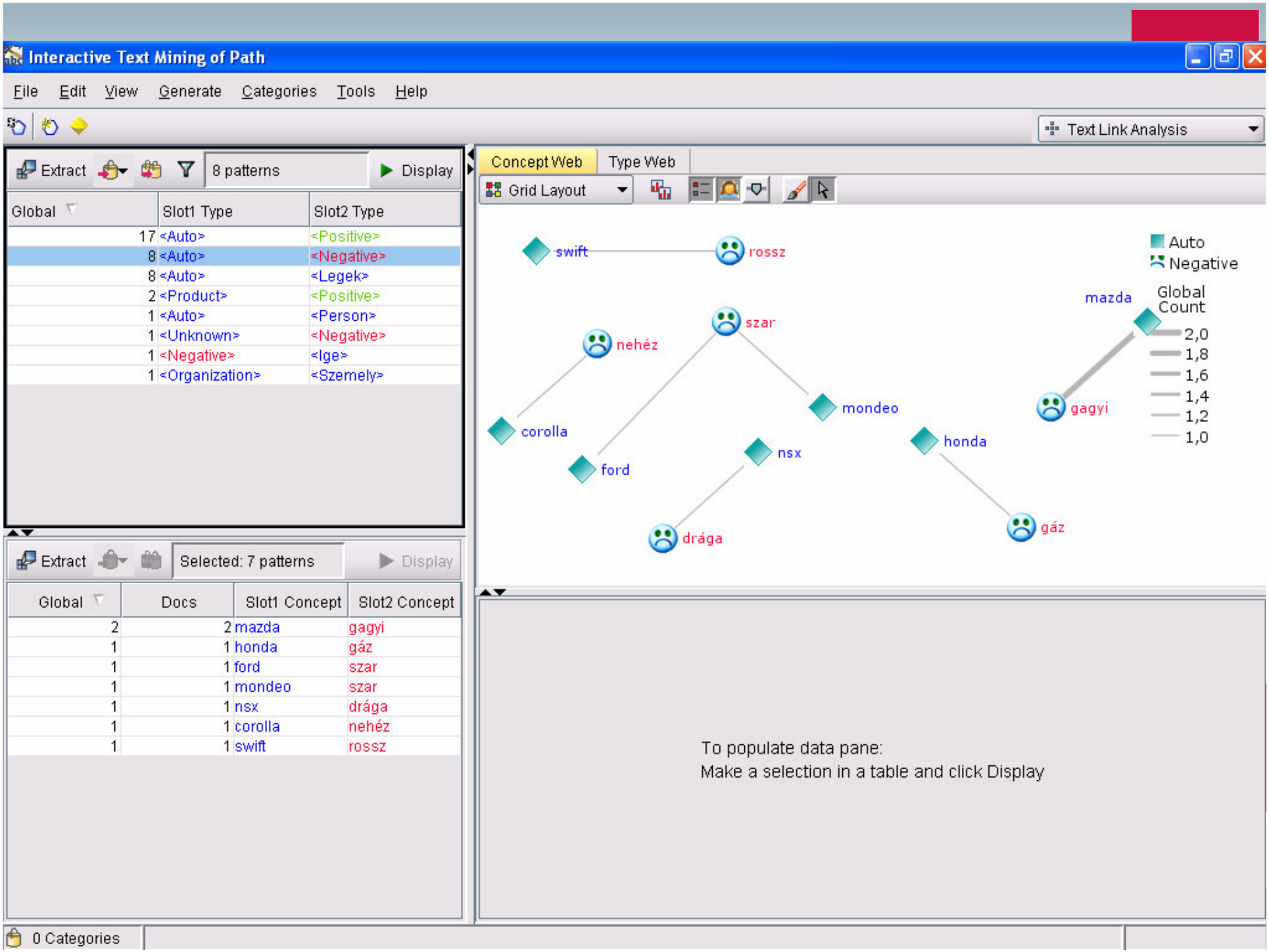


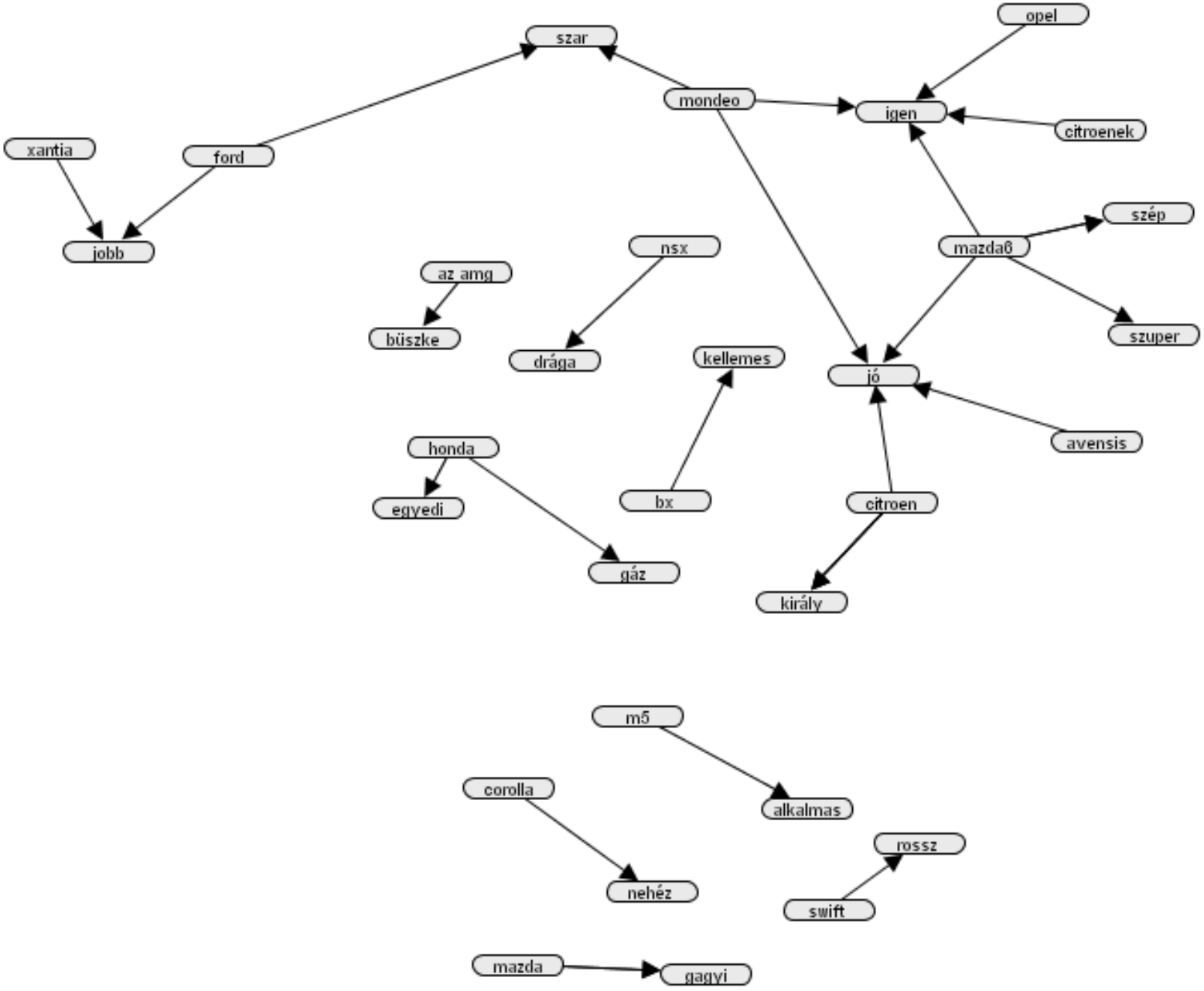


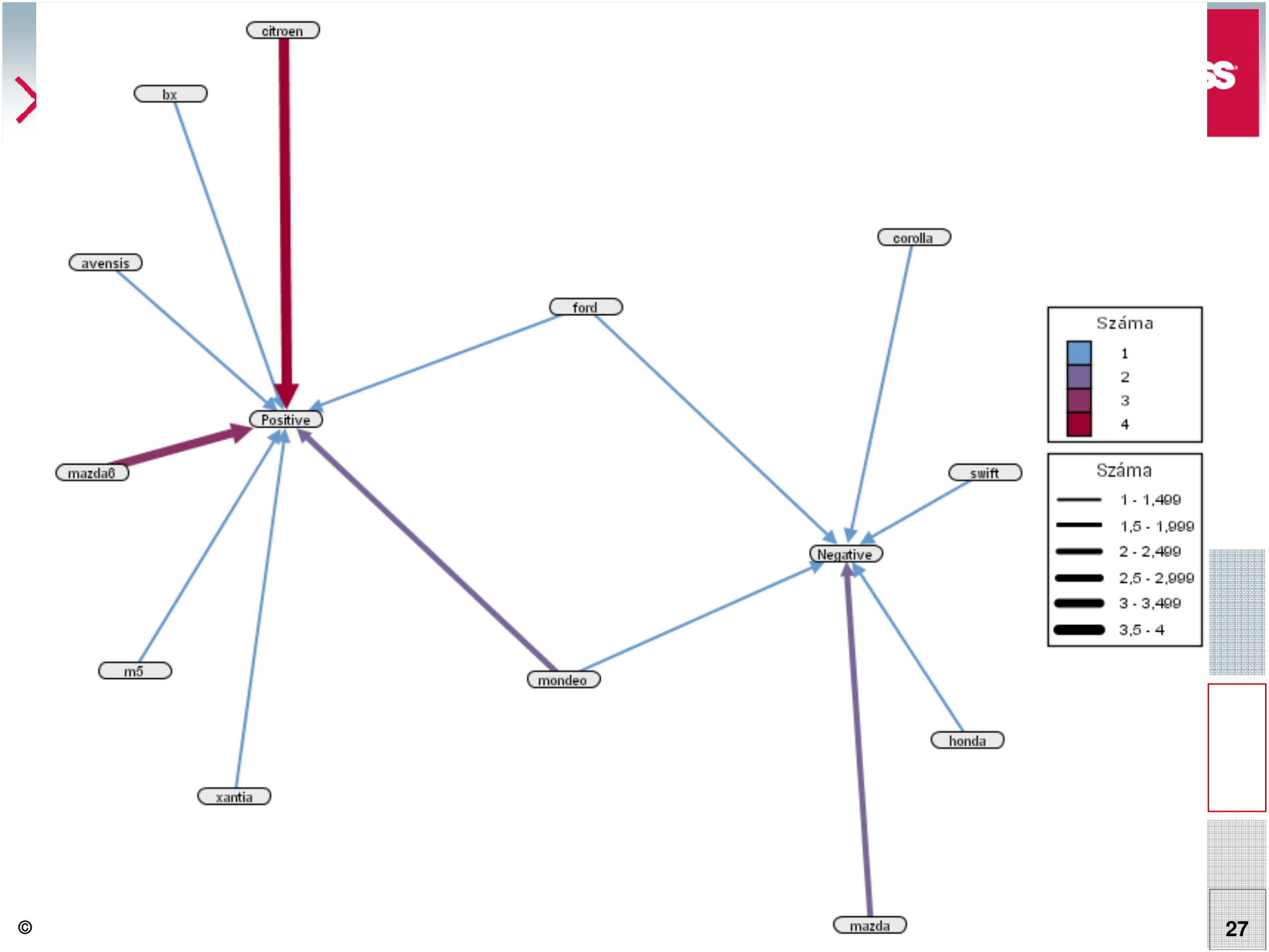
Magyar példák




Interactive Text Mining of Path
File Edit View Generate Categories Tools Help
Text Link Analysis
Extract
8 patterns
Display
Global
Slot1 Type
Slot2 Type
17 <Auto> <Positive>
8 <Auto> <Negative>
8 <Auto> <Legek>
2 <Product> <Positive>
1 <Auto> <Person>
1 <Unknown> <Negative>
1 <Negative> <Ige>
1 <Organization> <Szemely>
Concept Web
Type Web
Grid Layout
citroen
mazda6
szuper
mondeo
jó
avensis
király
büszke
az amg
szép
xantia
egyedi
ford
jobb
bx
honda
alkalmas
m5
kellemes
Auto
Positive
Global Count
3,0
2,5
2,0
1,5
1,0
Extract
Selected: 14 patterns
Display
Global
Docs
Slot1 Concept
Slot2 Concept
3 3 citroen király
2 2 mazda6 szép
1 1 xantia jobb
1 1 mondeo jó
1 1 az amg büszke
1 1 citroen jó
1 1 ford jobb
1 1 avensis jó
1 1 honda egyedi
1 1 mazda6 szuper
1 1 mazda6 jó
1 1 bx kellemes
1 1 mondeo jobb
1 1 m5 alkalmas
To populate data pane:
Make a selection in a table and click Display
0 Categories


Interactive Text Mining of Path
File Edit View Generate Categories Tools Help
Text Link Analysis
Extract
8 patterns
Display
Global
Slot1 Type
Slot2 Type
17 <Auto> <Positive>
8 <Auto> <Negative>
8 <Auto> <Legek>
2 <Product> <Positive>
1 <Auto> <Person>
1 <Unknown> <Negative>
1 <Negative> <Ige>
1 <Organization> <Szemely>
Concept Web
Type Web
Grid Layout
swift
rossz
szar
nehéz
mondeo
corolla
ford
nsx
drága
honda
gáz
mazda
gagyi
Auto
Negative
Global Count
2,0
1,8
1,6
1,4
1,2
1,0
Extract
Selected: 7 patterns
Display
Global
Docs
Slot1 Concept
Slot2 Concept
2 2 mazda gagyi
1 1 honda gáz
1 1 ford szar
1 1 mondeo szar
1 1 nsx drága
1 1 corolla nehéz
1 1 swift rossz
To populate data pane:
Make a selection in a table and click Display
0 Categories


opel
szar
mondeo
igen
citroenek
xantia
ford
jobb
szép
mazdaß
nsx
az amg
büszke
drága
szuper
kellemes
jó
honda
avensis
egyedi
bx
citroen
gáz
király
m5
corolla
alkalmas
rossz
nehéz
swift
mazda
gagyi


citroen
bx
avensis
ford
corolla
Positive
mazda6
swift
Negative
m5
mondeo
honda
xantia
mazda
Száma
1
2
3
4
Száma
1 - 1,499
1,5 - 1,999
2 - 2,499
2,5 - 2,999
3 - 3,499
3,5 - 4

TMC 3.0 Advanced Features (5)
Link Analysis & Pattern Matcher

- The pattern matcher allows you to find relationships between concepts identified during the text extraction process.

  Examples:
  - Bioinformatics: Gene1 <inhibits> Gene2
  - CRM: Customer1 <unhappy> handset
  - Homeland security: person <member of> organization

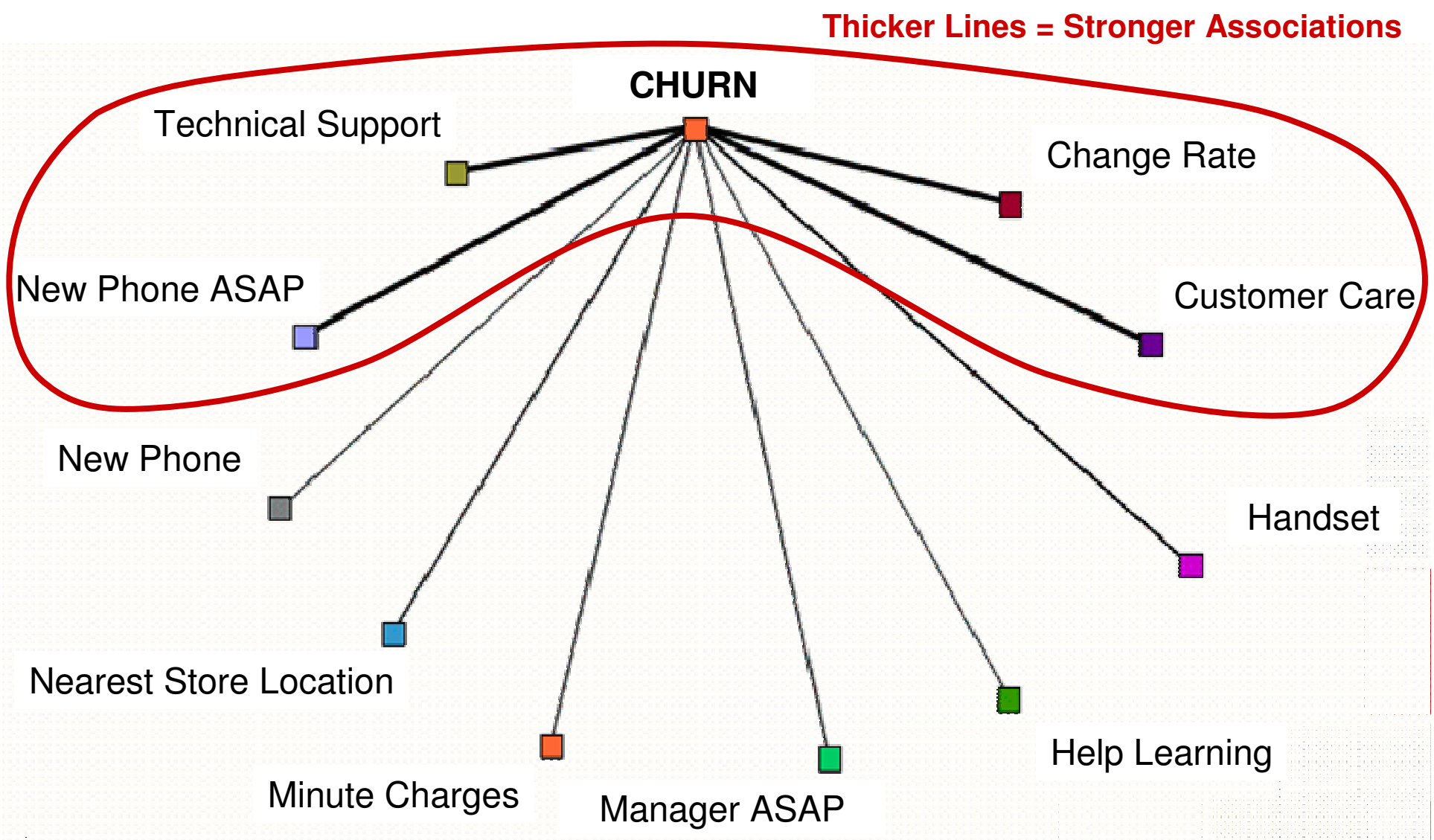
- TMC 3.0 gives you access to the pattern matcher via the Text Link Analysis node
  - CEMI node – load through CEMI dialogue

- Projects involving the pattern matcher should involve the Text Mining Task Force (contact Olivier Jouve)

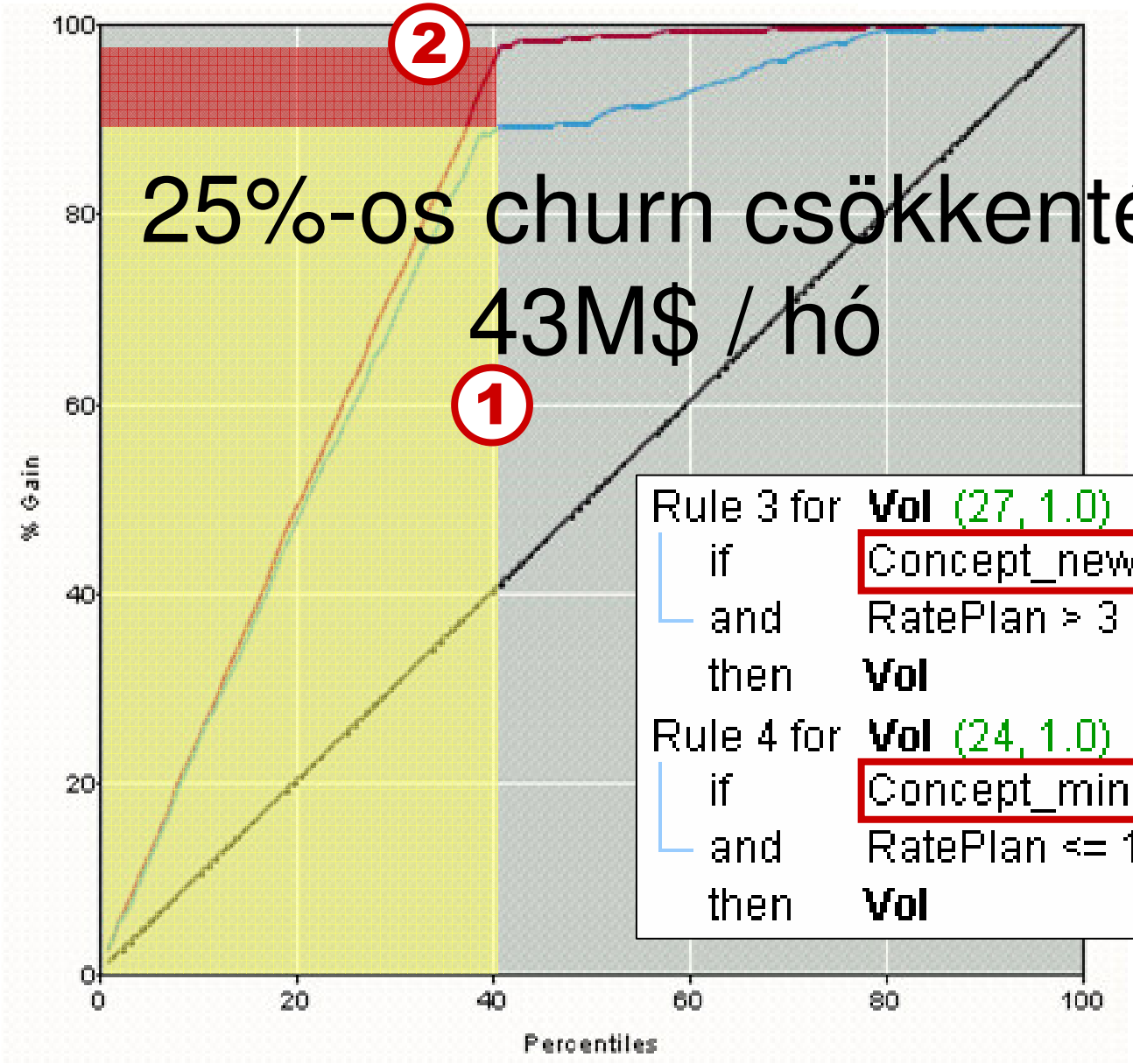
Mire jó a szöveganalitika?

- Security
- Attitűd (vélemény) azonosítás (pl. ügyfélszolgálat, piackutatás)
- Pharma
  - vényadatok
  - digitalizált levéltárak

Churnmodell javítása



CRM - telco



25%-os churn csökkentés

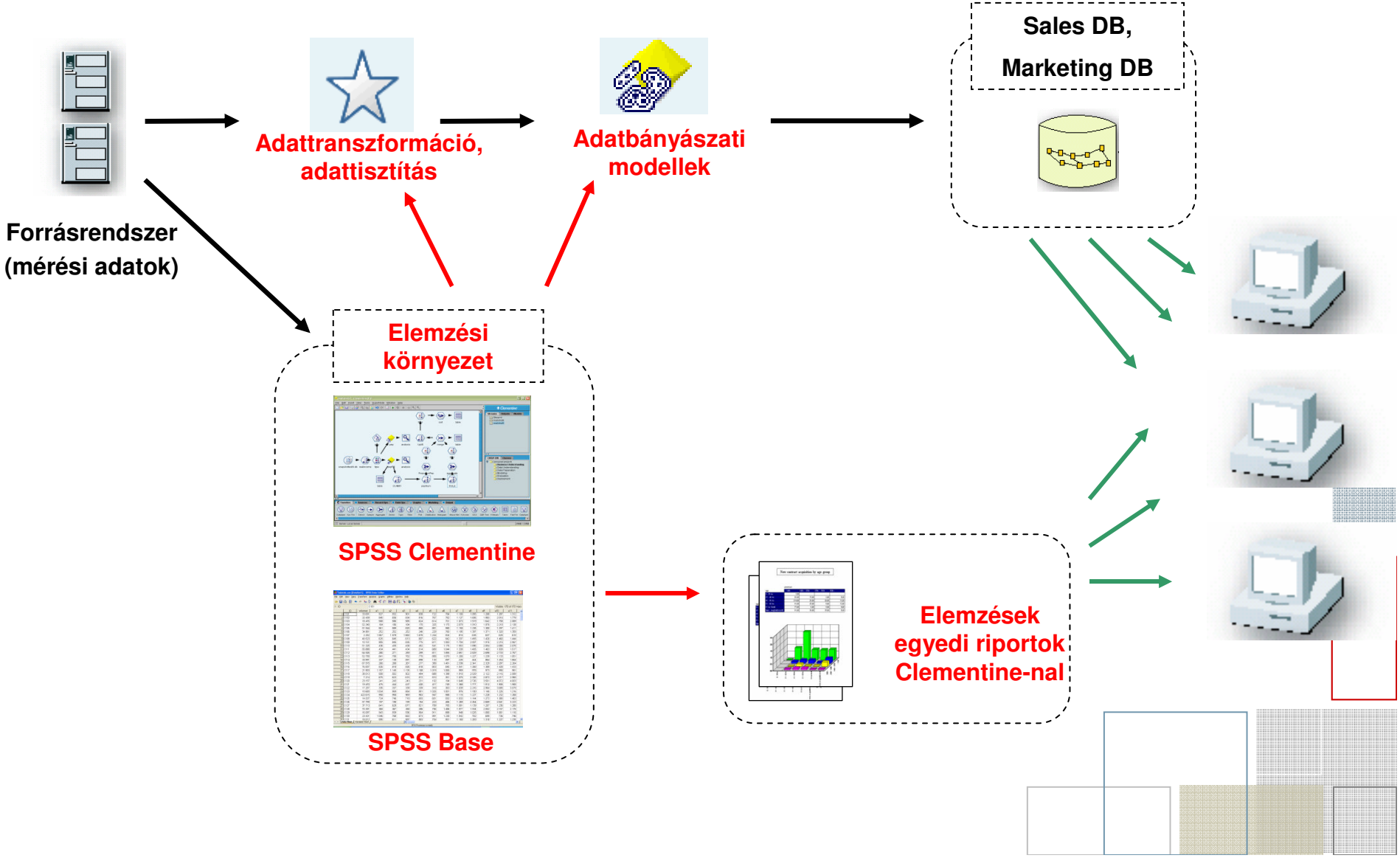
43M$ / hó

1. Hagyományos churn modell
2. Text mining eredményekkel

Rule 3 for **Vol** (27, 1.0)
- if Concept_new_phone_asap = 1
- and RatePlan > 3
- then **Vol**

Rule 4 for **Vol** (24, 1.0)
- if Concept_minute_charges = 1
- and RatePlan <= 1
- then **Vol**

Alkalmazás


Forrásrendszer
(mérési adatok)
Adattranszformáció,
adattisztítás
Adatbányászati
modellek
Sales DB,
Marketing DB
Elemzési
környezet
SPSS Clementine
SPSS Base
Elemzések
egyedi riportok
Clementine-nal

Kérdések?

Körmendi György
gykormendi@spss.hu
30-400-1854