Magyar Tudományos Akadémia
Nyelvtudományi Intézet

Intelligens elektronikus szótár és lexikai adatbázis

IHM-ITEM 48/2002

Oravecz Csaba
MTA Nyelvtudományi Intézet
Korpusznyelvészeti osztály
oravecz@nytud.hu

Bevezetés

- a jelenlegi helyzet: nincs az informatikai rendszerekben alkalmazható, megfelelő kifejtettségű nyelvi információt tartalmazó elektronikus szótár, lexikai adatbázis

- cél: nagyméretű lexikai adatbázis (LAB) kifejlesztése

- kiindulópont:
  - tartalmi: Magyar értelmező kéziszótár (ÉKsz.) átdolgozott változat; kb. 70.000 címszó
  - technológiai: CONCEDE (Consortium for Central European Dictionary Encoding) COPERNICUS projekt

- eszköz: CONCEDE technológia + humán erőforrás

Szótárak és lexikai adatbázisok

- könyv alakú szótár
  - a nyelvet értő emberi olvasásra készült
  - egyedi szerkezeti felépítésű
  - szerkezeti elemek jelölése tipográfiai jegyekkel
  - erősen tömörített, nehezen formalizálható információ
  - szerkezeti, tartalmi, tipográfiai hibákkal

Szótárak és lexikai adatbázisok

- géppel olvasható szótár
  - lényegében papírszótár vmilyen elektronikus formátumban
  - nagyszámú szerkezeti elem egyértelmű interpretáció nélkül
- lexikai adatbázis (LAB)
  - sztenderdizált, jól definiált szerkezet
  - elkülönített szerkezeti és tartalmi elemek
  - nyelvtechnológiai szempontból releváns információ
  - kevés elem, jól definiált interpretációval
  - előállítás: tartalmi és szerkezeti explikációval (up-translation)

Az ÉKsz könyvváltozata

hegedül tn és ts ige **1.** Hegedűn játszik (vmit).
2. *vál* <Tücsök> ciripel.

- felépítés: címszó szótári alak; bevezető rész; értelmező és szemléltető rész
- információ feldolgozása, "kinyerése": a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
- a címszó lineáris feldolgozása során végzett számos implicit "nyelvfeldolgozó", értelmező művelet (pl. tömörítések feloldása)

Az ÉKsz könyvváltozata

hegedül tn és ts ige **1.** Hegedűn játszik (vmit).
2. *vál* <Tücsök> ciripel.

- felépítés: címszó szótári alak; bevezető rész; értelmező és szemléltető rész
- információ feldolgozása, "kinyerése": a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
- a címszó lineáris feldolgozása során végzett számos implicit "nyelvfeldolgozó", értelmező művelet (pl. tömörítések feloldása)

Az ÉKsz könyvváltozata

hegedül tn és ts ige **1.** Hegedűn játszik (vmit).
2. *vál* <Tücsök> ciripel.

- felépítés: címszó szótári alak; bevezető rész; értelmező és szemléltető rész
- információ feldolgozása, "kinyerése": a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
- a címszó lineáris feldolgozása során végzett számos implicit "nyelvfeldolgozó", értelmező művelet (pl. tömörítések feloldása)

Az ÉKsz könyvváltozata

hegedül tn és ts ige **1.** Hegedűn játszik (vmit).
2. *vál* <Tücsök> ciripel.

- felépítés: címszó szótári alak; bevezető rész; értelmező és szemléltető rész
- információ feldolgozása, "kinyerése": a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
- a címszó lineáris feldolgozása során végzett számos implicit "nyelvfeldolgozó", értelmező művelet (pl. tömörítések feloldása)

Az első elektronikus változat

Példa

```
\entry{\lemma{hegedül\qlemma}
\gramgrp{\subc{tn\qsubc} \pos{ige\qpos}\qgramgrp} \es{}
\gramgrp{\subc{ts\qsubc} \pos{ige\qpos}\qgramgrp}
\sense{\num{1.}
\defi{hegedűn játszik \hint{vmit\qhint}.\qdefi}
\qsense{1.}}
\sense{\num{2.}
\usg{vál\qusg}
\defi{\gloss{Tücsök\qgloss} ciripel.\qdefi}
\qsense{2.}}
\qentry}
```

Az első elektronikus változat

- félúton a papírszótár és a géppel olvasható szótár között
- kísérlet a címszó elemeinek és a köztük lévő hierarchikus viszonyoknak a leképezésére
- következetlen kódolás, nehezen azonosítható hibák
- gépi ellenőrzés nem alkalmazható

A géppel olvasható szótár

- az ÉKsz elektronikus változata, ezért ugyanabban a lineáris és tömörített formában hordozza az információt
- tipográfia még mindig elsődleges
- szabványos XML kódolású
  - szócikkelemek
  - elemszerkezet DTD által meghatározott
  - "megengedő" DTD, nagyszámú elem, nem egyértelmű interpretáció

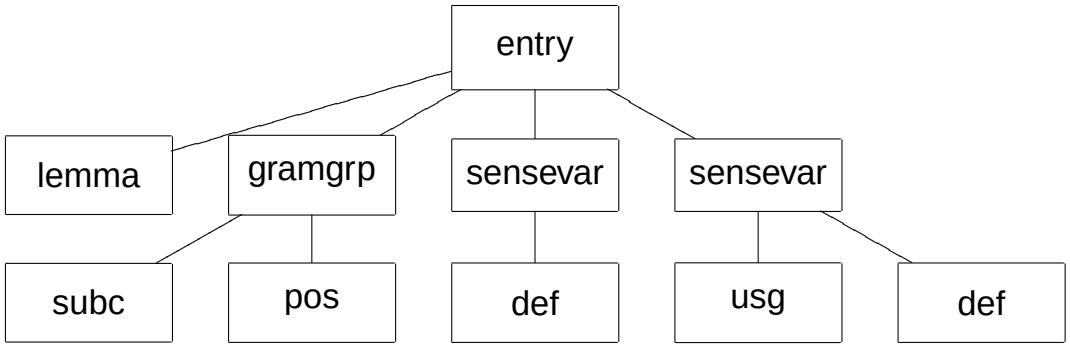
A kiinduló XML változat

Példa

```
<entry id="id-1300300">
 <lemma>hegedül</lemma>
 <gramgrp>
  <subc>tn <logi type="es"/> ts</subc>
  <pos>ige</pos>
 </gramgrp>
 <sensevar>
  <def>Hegedűn játszik <hint>vmit</hint>.</def>
 </sensevar>
 <sensevar>
  <usg>vál</usg>
  <def><gloss>Tücsök</gloss> ciripel.</def>
 </sensevar>
</entry>
```

Szócikkszerkezet

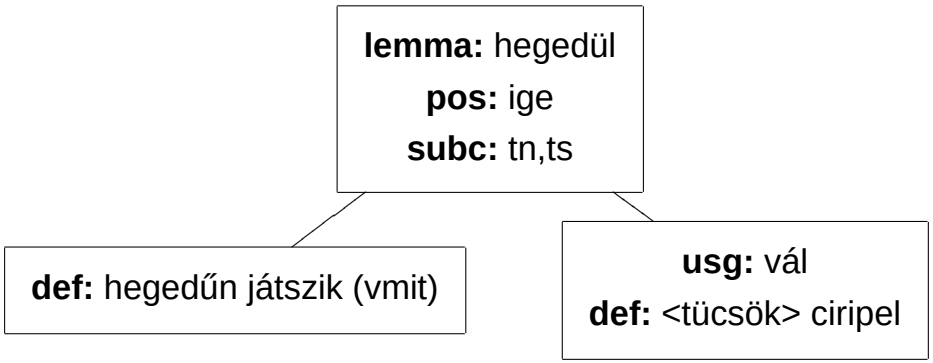
- DTD: kiterjesztett környezetfüggetlen nyelvtan (ECFG)
- leírt szerkezet egy fában ábrázolható



1. ábra. A példaszócikk elemágrajza.

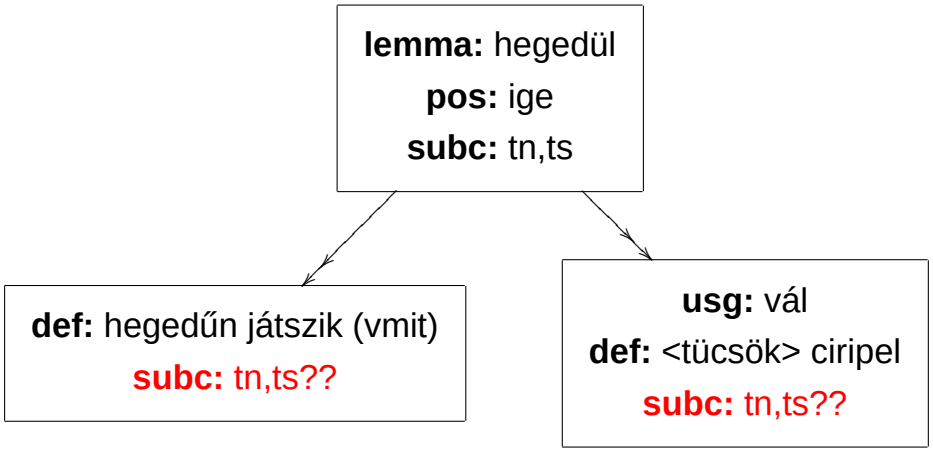
Szócikkszerkezet

- a szócikk emberi feldolgozásának gépi modellje (információkinyerés): fabejárás
- problémák:
  - milyen elemek vezetnek be új információs csomópontot?
  - hogyan "áramlik" az információ egyik csomópontról a másikra? (szülő csomópontról öröklődik $\Rightarrow$ SUBC??)



Szócikkszerkezet

- a szócikk emberi feldolgozásának gépi modellje (információkinyerés): fabejárás
- problémák:
  - milyen elemek vezetnek be új információs csomópontot?
  - hogyan "áramlik" az információ egyik csomópontról a másikra? (szülő csomópontról öröklődik $\Rightarrow$ SUBC??)



A lexikai adatbázis

- szerkezeti elem: információs csomópontot reprezentál (STRUC)

  értelmezés: diszjunktív; az adott címszó egy-egy elemi használati módját reprezentálja

- tartalmi elemek: információt hordoznak

  (jegy-érték reprezentációban: jegy = elem neve; érték = elem tartalma)

  értelmezés: konjunktív; egy szerkezeti elem közvetlen leszármazott tartalmi elemei által hordozott információ mind érvényes az adott csomópontban

A lexikai adatbázis

- információáramlás: öröklődés a csomópontok között
  - kumulatív: a csomópontok bejárása során az azonos jegyek értéke összeadódik (pl. USG?)
  - felülíró: az azonos jegyek közül az adott csomópontban a legközelebbi értéke érvényes

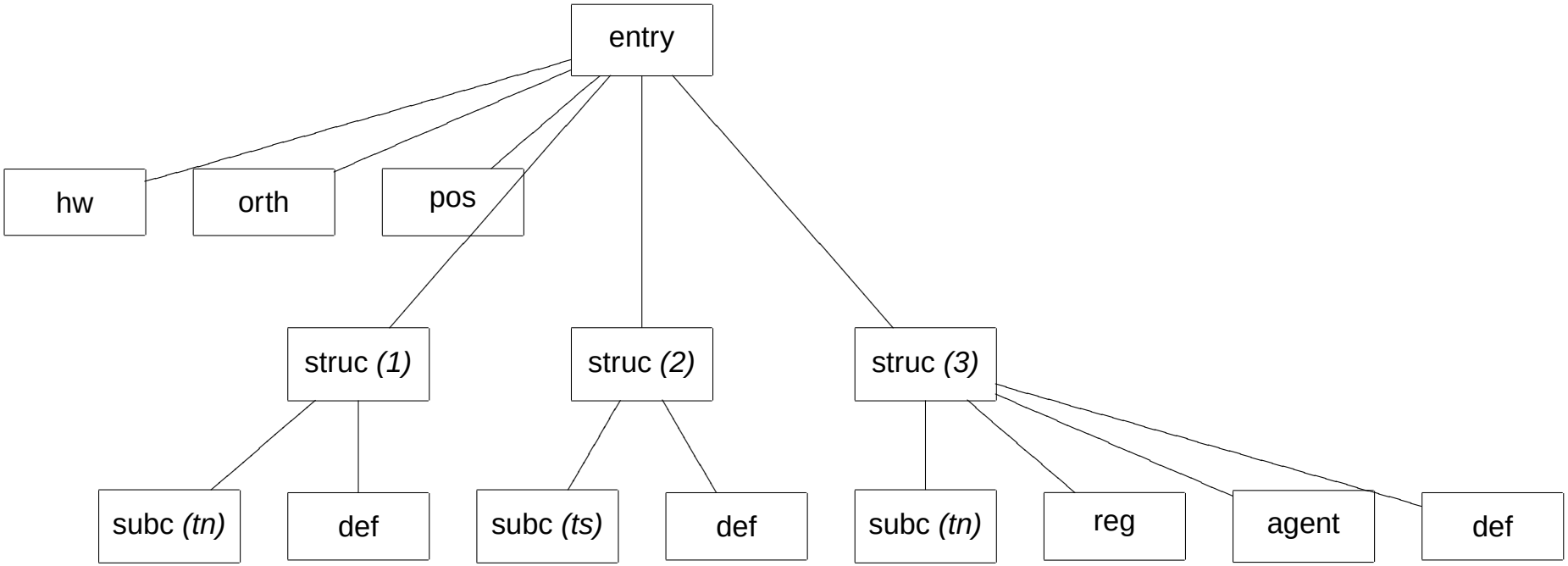
- tartalmi elemek meghatározása:
  - azon információtípusok, ahol a felülírás nem áll fenn, különböző elemekként jelennek meg (USG $\Rightarrow$ REG, GEO, DOMAIN stb.)
  - egymást felülíró információtípusok: azonos elem különböző tartalom (POS, SUBC stb)

Az adatbázis formátum

Példa

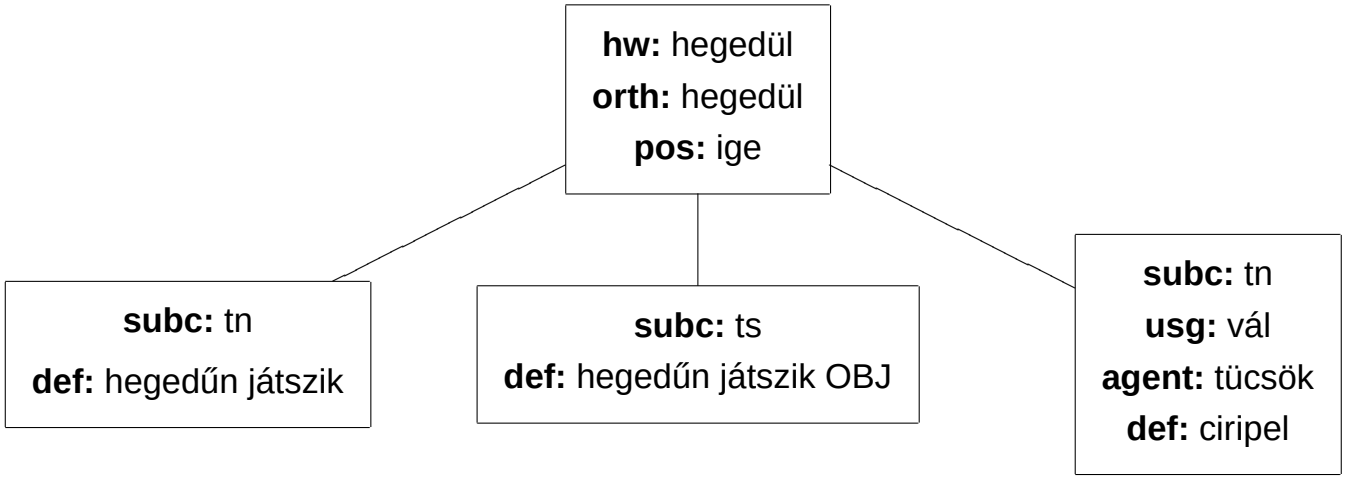
```
<entry id="hegedül.1">
 <hw>hegedül</hw><orth>hegedül</orth><pos>ige</pos>
 <struc id="hegedül.1.1" type="sense">
  <subc>tn</subc>
  <def>hegedűn játszik</def>
 </struc>
 <struc id="hegedül.1.2" type="sense">
  <subc>ts</subc>
  <def>hegedűn játszik <obj/></def>
 </struc>
 <struc id="hegedül.1.3" type="sense">
  <subc>tn</subc><reg>vál</reg><agent>tücsök</agent>
  <def>ciripel</def>
 </struc>
</entry>
```

Szócikkszerkezet a LAB-ban



2. ábra. A példaszócikk elemágrajza a LAB-ban.

Szócikkszerkezet a LAB-ban



3. ábra. A fabejárás során kiolvasható információ a LAB-ban.

Információkinyerés a LAB-ból

- a szócikkszerkezet minden egyes csomópontjához egyértelmű és kimerítő információ tartozik

Információkinyerés a LAB-ból

- a szócikkszerkezet minden egyes csomópontjához egyértelmű és kimerítő információ tartozik

1. az adott csomópontnál van megadva

Információkinyerés a LAB-ból

- a szócikkszerkezet minden egyes csomópontjához egyértelmű és kimerítő információ tartozik

1. az adott csomópontnál van megadva
2. öröklődik magasabb csomópontból

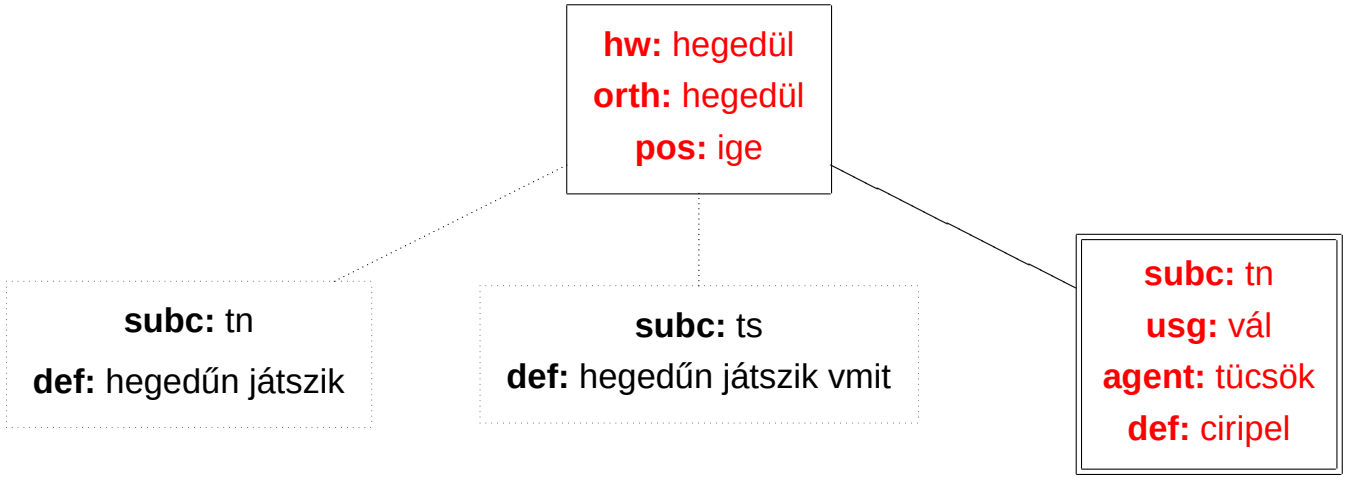
Információkinyerés a LAB-ból

- a szócikkszerkezet minden egyes csomópontjához egyértelmű és kimerítő információ tartozik

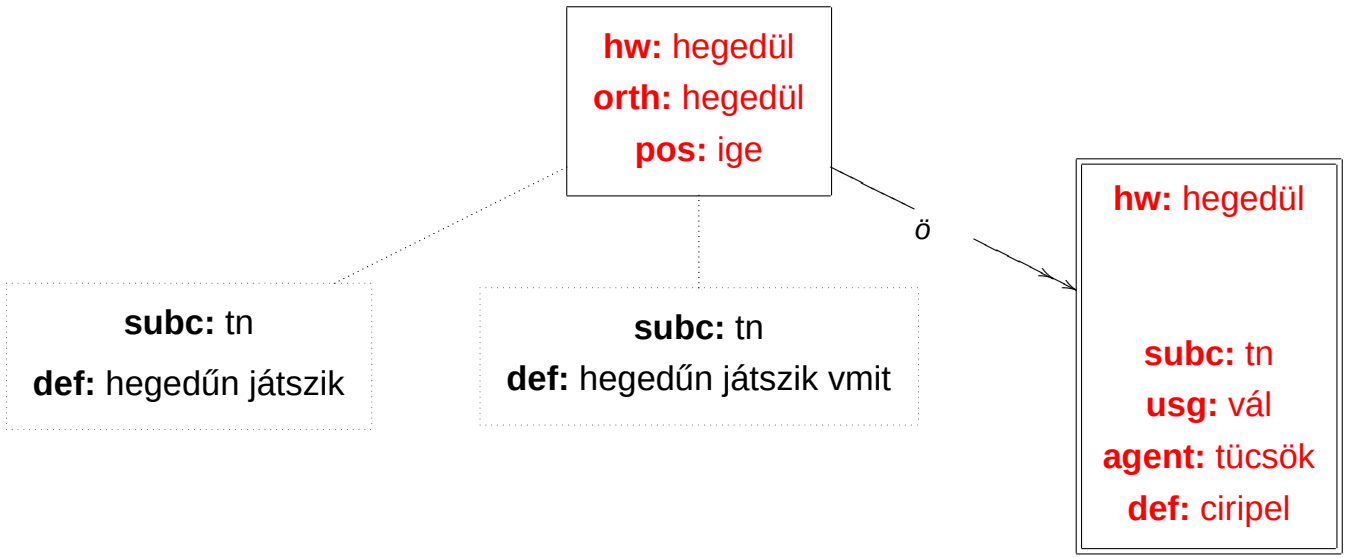
  1. az adott csomópontnál van megadva
  2. öröklődik magasabb csomópontból

- minél alacsonyabb szintű csomópont a fában, annál specifikusabb információt tartalmaz az adott címszó egy elemi használatáról

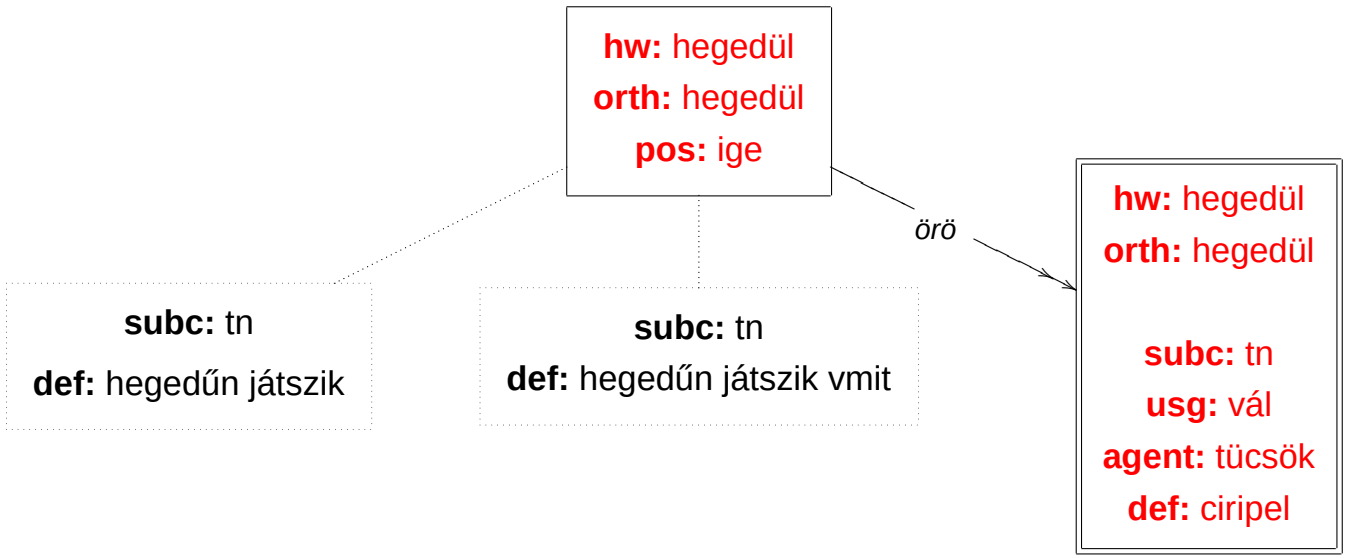
Információkinyerés a LAB-ból



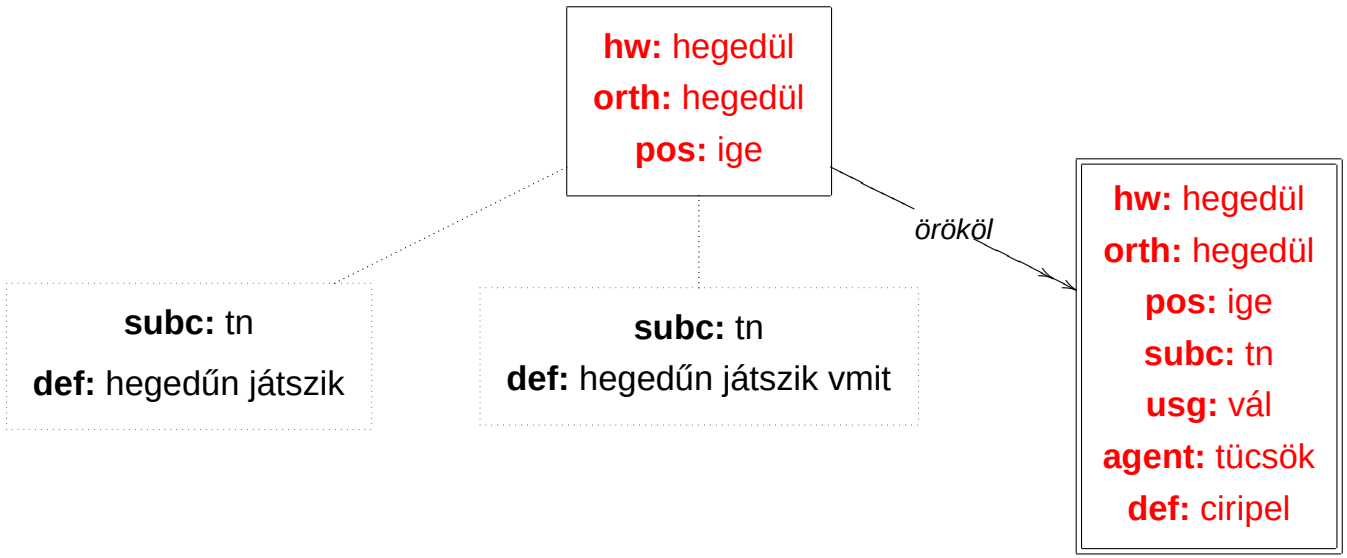
Információkinyerés a LAB-ból



Információkinyerés a LAB-ból



Információkinyerés a LAB-ból



Feladatok a LAB kialakítása során

- tipográfiai szempontú kódolás átalakítása (különböző elemek azonos tipográfiával $\Rightarrow$ következetlen kódolás)
- szótári tömörítés feloldása
  - szerkezeti (v.ö. "hegedül")
  - tartalmi

bronz . . . **1.** Réznek és ónnak az ötvözete.
. . . **2.** *biz* Ebből készült (mű)tárgy, kül. érem.

ad . . . *Vmire v. vminek adja magát v. a fejét*: vmire szánja, ill. vminek átengedi magát.

Feladatok a LAB kialakítása során

- nem releváns információ törlése (központozás, nyelvhelyességi "útmutató"), hibajavítás
- validáció
  - szerkezeti: XML validáló elemző
  - tartalmi: manuális ellenőrzés kiválasztott elemeken

A projekt várható eredménye

- nagyméretű, szabványos formátumú lexikai adatbázis
  - explicit, jól definiált, géppel értelmezhető reprezentáció
  - számítógépes nyelvfeldolgozó alkalmazások alapja
- hálózati lekérdező felület
  - strukturált formában kinyerhető információ
  - többszempontú lekérdezési lehetőség

VÉGE