Nyelvtechnológia - nyelvészeknek

A korpusznyelvészettől a nyelvtechnológiáig

Váradi Tamás

MTA Nyelvtudományi Intézet
varadi@nytud.hu

A 2006. november 30-án a Nyelvtudományi Intézetben tartott előadás bővített változata

Vázlat

Vázlat

Vázlat

Vázlat

Vázlat

Vázlat

Motiváció

Nyelvtechnológia – nyelvészeknek

- nyelvészeknek - azaz nem informatikusoknak
- a nyelvtechnológiát az informatika hívta létre
- az informatikában egyértelműen igazolta magát

- Vezérmotívum:
  - Mennyi benne a nyelvészet?
  - Mennyiben releváns a nyelvészet számára?
  - Mit nyújt a nyelvészeknek?

Vázlat

A kihívás

A nyelvet a számítógép számára érthetővé tenni

- szövegek, szótárak, nyelvtanok — emberek készítik embereknek
- értésükhöz, alkalmazásukhoz nyelvi és világismeret kell
- a számítógép számára mindezt expliciten meg kell adni
- az igazi generatív vállalkozás

Szemléleti különbség

Középpontban a beszéd (*parole*)

- Feladat: a nyelv visszafejtése (*reverse engineering*)
- nem „csak” a nyelv, hanem a nyelvhasználat (*performancia*)
- adatok és eljárások, algoritmusok — működő rendszer
- végső soron az emberi beszédértés, beszédalkotás szimulálása

A robusztusság alapkövetelmény

- a gond nem az adatok tömege, hanem „fésületlenségük”
- a szőnyeg alá söprés nem megy

Szemléleti különbség

Középpontban a beszéd (*parole*)

- Feladat: a nyelv visszafejtése (*reverse engineering*)
- nem „csak” a nyelv, hanem a nyelvhasználat (*performancia*)
- adatok és eljárások, algoritmusok — működő rendszer
- végső soron az emberi beszédértés, beszédalkotás szimulálása

A robusztusság alapkövetelmény

- a gond nem az adatok tömege, hanem „fésületlenségük”
- a szönyeg alá söprés nem megy

Szemléleti különbség

Rapid megoldás mindenek felett

- Terjedő paradigma: statisztikai nyelvi modellezés
  `http://nlp.stanford.edu/links/statnlp.html`
  - nyelvfüggetlen eljárás
  - kiinduló adathalmaz ún. tanuló korpusz
  - gépi tanulás
    `http://en.wikipedia.org/wiki/Machine_learning`

Vázlat

Ki a korpusznyelvész?

Aki korpuszokat alkalmaz?

- Korpuszt használni = független, külső adatokat alkalmazni
- a korpuszok használata egyre jobban beépül a nyelvészeti gyakorlatba
- ettől még ki-ki megmarad francia, finnugor stb. nyelvésznek

Aki korpuszokat készít!

- A korpuszok készítése önálló szakma
  - A korpuszok megtervezése
  - összeállítása
  - nyelvi elemzése
  - működtetése
  - karbantartása
- a korpusznyelvészet feladata

Ki a korpusznyelvész?

Aki korpuszokat alkalmaz?

- Korpuszt használni = független, külső adatokat alkalmazni
- a korpuszok használata egyre jobban beépül a nyelvészeti gyakorlatba
- ettől még ki-ki megmarad francia, finnugor stb. nyelvésznek

Aki korpuszokat készít!

- A korpuszok készítése önálló szakma
  - A korpuszok megtervezése
  - összeállítása
  - nyelvi elemzése
  - működtetése
  - karbantartása
- a korpusznyelvészet feladata

Nem a Web a legjobb korpusz?

Miért nem?

- Teljesen bizonytalan eredetű (akár nem anyanyelvi) szövegek
- Méretét is legfeljebb becsülni lehet

Miért érdekes mégis?

- Elképesztő tömegű szöveg
- Rendkivül gyorsan nő
- A „legdemokratikusabb” médium: a beszélők minden eddiginél szélesebb körét reprezentálja
- Bizonyos célokra így is jó, ahogy van (ld. a köv. táblázat)

Nem a Web a legjobb korpusz?

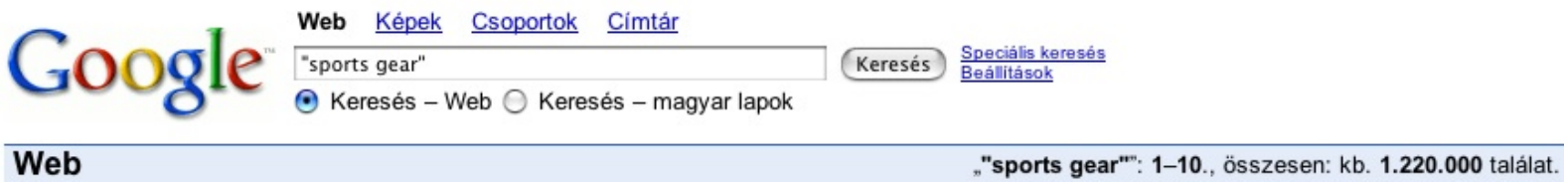
Miért nem?

- Teljesen bizonytalan eredetű (akár nem anyanyelvi) szövegek
- Méretét is legfeljebb becsülni lehet

Miért érdekes mégis?

- Elképesztő tömegű szöveg
- Rendkivül gyorsan nő
- A „legdemokratikusabb” médium: a beszélők minden eddiginél szélesebb körét reprezentálja
- Bizonyos célokra így is jó, ahogy van (ld. a köv. táblázat)

Nem a Web a legjobb korpusz?



sports gear	1.220.000
sporting gear	179.000
sports equipment	1.480.000
sporting equipment	1.070.000
sports geer	73
sporting geer	2

A „sportszer” szó lehetséges angol megfeleléseinek gyakorisága

Nem a Web a legjobb korpusz? (folyt.)

Konklúzió

- Gyors, elnagyolt mintavétel
- Bizonyos durva különbségekre jól használható
- Az elképesztően nagy és rohamosan növekvő méret páratlan előny
- Meg kell tanulni kihasználni az előnyeit

Mitől korpusz egy halom szöveg?

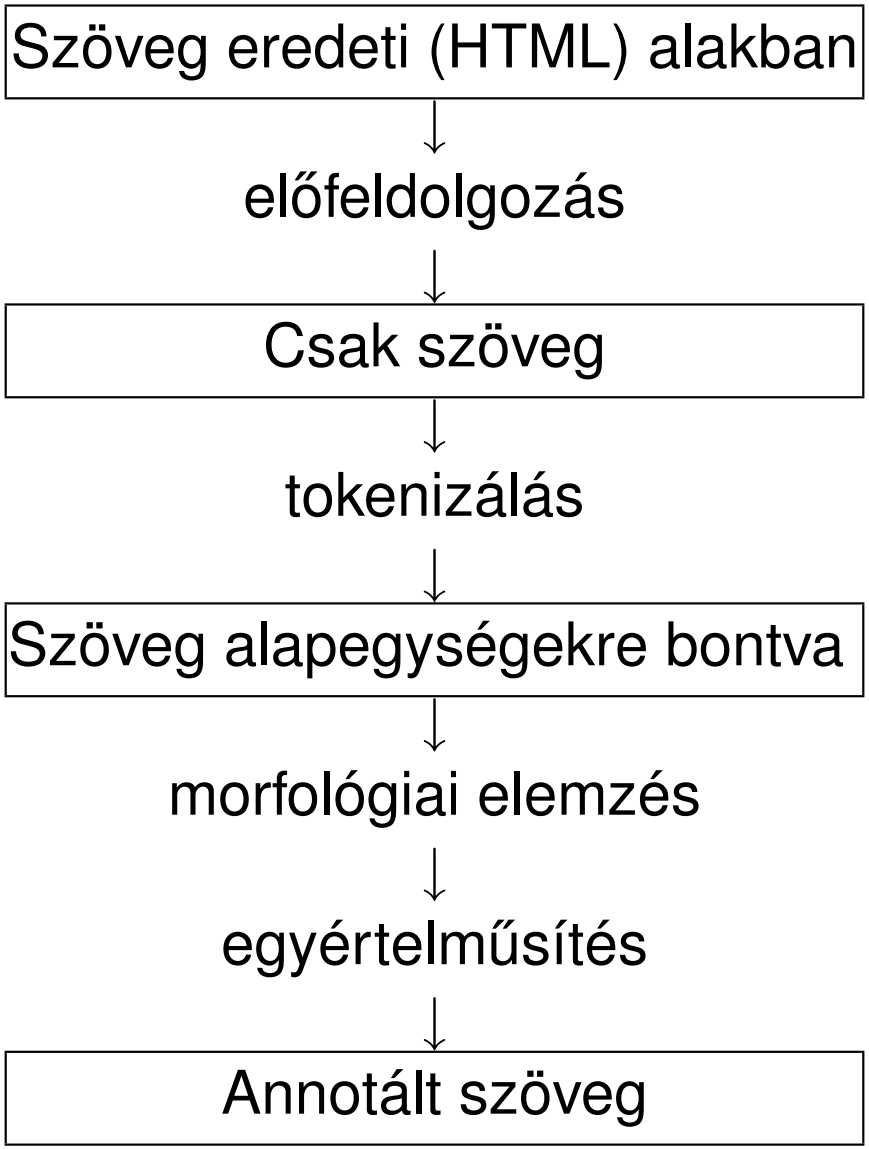
Korpusz <=> szövegarchívum

- Korpusz:
  - egységes elvek szerinti válogatás
  - egységes kódolás

Mitől korpusz egy halom szöveg?

Korpusz <=> szövegarchívum

- Korpusz:
  - egységes elvek szerinti válogatás
  - egységes kódolás


Szöveg eredeti (HTML) alakban
előfeldolgozás
Csak szöveg
tokenizálás
Szöveg alapegységekre bontva
morfológiai elemzés
egyértelműsítés
Annotált szöveg

MNSZ részlet 1

```
<?xml version="1.0" encoding="iso-8859-2" standalone="ye
<text>
 <!--beginning of orig-->
 <!-- Digitalis Archivum ## /home2/projects/sulinet_ihm2
<div id="lit-dia-Bella_Istvan___Hetedik_kavics___1975.cl
<head>
 <s>
  <title type="konyvcim">
   <w LEMMA="Hetedik" CAT="Num" NOM>Hetedik</w>
   <w LEMMA="kavics" CAT="N" NOM>kavics</w>
  </title>
 </s>
</head>
```

MNSZ részlet 2

```
<poem>
<lg>
<l>
<w LEMMA=szanaszét" CAT="Adv">Szanaszét</w>
<w LEMMA="széled" CAT="V" e M 3>széledt</w>
<w LEMMA="ujj" CAT="N" e 1 INS PS i>ujjaimmal</w>
</l>
<l>
<w LEMMA="elveszett" CAT="MIB" NOM>elveszett</w>
<w LEMMA="koponya" CAT="N" e 1 NOM PS>koponyám</w>
<w LEMMA="most" CAT="Adv">most</w>
<w LEMMA="megkeres" CAT="V" e 1 T Pre>megkeresem</w>
<c lemma="," msd="WPUNCT" ctag="WPUNCT">,</c>
</l>
</lg>
```

Korpusznyelvészet pro és kontra

Mellette

- tényleges nyelvhasználat
- objektív adatok
- új dimenzió: gyakoriság
- sokaság (nagy számok törvénye)

Ellene

- a mintavétel módszertana kétséges
- a nyelv fogalma aluldefiniált
- nem ad számot a potenciális alakokról
- adatok nem tiszták (performancia) hibák

Konklúzió

- a korpusz a nyelvhasználat lenyomata - nem nyújtja közvetlen a nyelvi rendszert
- ugyanúgy aluldefiniált mint maga a teljes nyelvhasználat

Korpusznyelvészet pro és kontra

Mellette

- tényleges nyelvhasználat
- objektív adatok
- új dimenzió: gyakoriság
- sokaság (nagy számok törvénye)

Ellene

- a mintavétel módszertana kétséges
- a nyelv fogalma aluldefiniált
- nem ad számot a potenciális alakokról
- adatok nem tiszták (performancia) hibák

Konklúzió

- a korpusz a nyelvhasználat lenyomata - nem nyújtja közvetlen a nyelvi rendszert
- ugyanúgy aluldefiniált mint maga a teljes nyelvhasználat

Korpusznyelvészet pro és kontra

Mellette

- tényleges nyelvhasználat
- objektív adatok
- új dimenzió: gyakoriság
- sokaság (nagy számok törvénye)

Ellene

- a mintavétel módszertana kétséges
- a nyelv fogalma aluldefiniált
- nem ad számot a potenciális alakokról
- adatok nem tiszták (performancia) hibák

Konklúzió

- a korpusz a nyelvhasználat lenyomata - nem nyújtja közvetlen a nyelvi rendszert
- ugyanúgy aluldefiniált mint maga a teljes nyelvhasználat

Vázlat

Néhány alapkérdés

A vizsgálandó adatok véges, zárt univerzumot alkotnak

- pl. az 2006. okt. 23-án elhangzott összes rendőrségi rádióadás
  - kimerítően lejegyezhető
  - a korpusz tartalmazza az ún. cél populációt

A vizsgálandó adatok véges, de túl nagy univerzumot alkotnak

- az okt. 23-án elhangzott vagy leírt összes magyar megnyilatkozás
  - elvileg véges, nagysága megbecsülhető
  - gyakorlatilag rögzíthetetlen
  - a digitális kultúra terjedtével írásos része egyre nagyobb mértékben elérhető elektronikusan
  - a korpusz statisztikai minta

Néhány alapkérdés

A vizsgálandó adatok véges, zárt univerzumot alkotnak

- pl. az 2006. okt. 23-án elhangzott összes rendőrségi rádióadás
  - kimerítően lejegyezhető
  - a korpusz tartalmazza az ún. cél populációt

A vizsgálandó adatok véges, de túl nagy univerzumot alkotnak

- az okt. 23-án elhangzott vagy leírt összes magyar megnyilatkozás
  - elvileg véges, nagysága megbecsülhető
  - gyakorlatilag rögzíthetetlen
  - a digitális kultúra terjedtével írásos része egyre nagyobb mértékben elérhető elektronikusan
  - a korpusz statisztikai minta

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
- a nyelvi változatokra

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
  - demográfiai mintavétel
  - vannak független adatok a beszélők csoportváltozóiról
- a nyelvi változatokra
  - nem ismerjük az egyes nyelvi változatok arányait a teljes nyelvhasználatan

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
  - demográfiai mintavétel
  - vannak független adatok a beszélők csoportváltozóiról
- a nyelvi változatokra
  - nem ismerjük az egyes nyelvi változatok arányait a teljes nyelvhasználatan

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
  - demográfiai mintavétel
  - vannak független adatok a beszélők csoportváltozóiról
- a nyelvi változatokra
  - nem ismerjük az egyes nyelvi változatok arányait a teljes nyelvhasználatan

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
  - demográfiai mintavétel
  - vannak független adatok a beszélők csoportváltozóiról
- a nyelvi változatokra
  - nem ismerjük az egyes nyelvi változatok arányait a teljes nyelvhasználatan

A korpusz mint minta

Mire legyen reprezentatív a korpusz?

- a beszélőkre
  - demográfiai mintavétel
  - vannak független adatok a beszélők csoportváltozóiról
- a nyelvi változatokra
  - nem ismerjük az egyes nyelvi változatok arányait a teljes nyelvhasználatan

A mérhető adat

A sokaság szerepe

- a korpusz megszámolhatóvá teszi az adatokat
- a nagy méret kiegyenlítő szerepet játszik
- ugyanakkor szinte kizárja a 100%-os pontosságot/adattisztaságot

A gyakoriság

- új dimenziót nyit a nyelvelemzésben
- függvénye a korpusz összetételének és méretének (minta arányos-e a teljességgel?)
- szerepe a nyelvi kompetenciában növekvő mértékben elismert
- az emberi nyelvfeldolgozás modellezésében fontos szerep

A mérhető adat

A sokaság szerepe

- a korpusz megszámolhatóvá teszi az adatokat
- a nagy méret kiegyenlítő szerepet játszik
- ugyanakkor szinte kizárja a 100%-os pontosságot/adattisztaságot

A gyakoriság

- új dimenziót nyit a nyelvelemzésben
- függvénye a korpusz összetételének és méretének (minta arányos-e a teljességgel?)
- szerepe a nyelvi kompetenciában növekvő mértékben elismert
- az emberi nyelvfeldolgozás modellezésében fontos szerep

Vázlat

Korpusznyelvészet és szociolingvisztika

Kezdetben (LOB, BROWN korpusz)

- hangsúly a nyelvhasználati változatokon
- forrásokról sok, részletes adat, szerzőkről nagyon kevés

Manapság (mega- és giga korpuszok)

- hangsúly egyértelműen az adatmennyiségen
- statisztikai nyelvfeldolgozás céljaira

Hiánycikk: homogén beszédközösséget megörökítő korpusz)

- demográfiailag – szociolingvisztikailag érvényes (pl. terepmunkából származó) korpusz
- Van: Labov gyűjtése
  `http://projects.ldc.upenn.edu/DASL/SLX/`
- Jön: BUSZI

Korpusznyelvészet és szociolingvisztika

Kezdetben (LOB, BROWN korpusz)

- hangsúly a nyelvhasználati változatokon
- forrásokról sok, részletes adat, szerzőkről nagyon kevés

Manapság (mega- és giga korpuszok)

- hangsúly egyértelműen az adatmennyiségen
- statisztikai nyelvfeldolgozás céljaira

▸ LDC

Hiánycikk: homogén beszédközösséget megörökítő korpusz

- demográfiailag – szociolingvisztikailag érvényes (pl. terepmunkából származó) korpusz
- Van: Labov gyűjtése `http://projects.ldc.upenn.edu/DASL/SLX/`
- Jön: BUSZI

Korpusznyelvészet és szociolingvisztika

Kezdetben (LOB, BROWN korpusz)

- hangsúly a nyelvhasználati változatokon
- forrásokról sok, részletes adat, szerzőkről nagyon kevés

Manapság (mega- és giga korpuszok)

- hangsúly egyértelműen az adatmennyiségen
- statisztikai nyelvfeldolgozás céljaira

Hiánycikk: homogén beszédközösséget megörökítő korpusz)

- demográfiailag – szociolingvisztikailag érvényes (pl. terepmunkából származó) korpusz
- Van: Labov gyűjtése
  `http://projects.ldc.upenn.edu/DASL/SLX/`
- Jön: BUSZI

Ritka kivételek

British National Corpus (BNC)

- www.natcorp.ox.ac.uk
- 10 %-nyi (10 m szó!) hanganyag demográfiai mintavétellel
- adatközlőkről gondos szociológiai nyilvántartás

International Corpus of English (ICE-GB)

- www.ucl.ac.uk/english-usage/projects/ice-gb
- 500 szöveg (1 m szó) nagyobb része, 300(!) hanganyag
- minden mondat szintaktikai szerkezete kézzel annotálva
- szintaktikai ágrajz és hallható hang
- fejlett keresési lehetőség a szintaktikai faszerkezetben

Ritka kivételek

British National Corpus (BNC)

- www.natcorp.ox.ac.uk
- 10 %-nyi (10 m szó!) hanganyag demográfiai mintavétellel
- adatközlőkről gondos szociológiai nyilvántartás

International Corpus of English (ICE-GB)

- www.ucl.ac.uk/english-usage/projects/ice-gb
- 500 szöveg (1 m szó) nagyobb része, 300(!) hanganyag
- minden mondat szintaktikai szerkezete kézzel annotálva
- szintaktikai ágrajz és hallható hang
- fejlett keresési lehetőség a szintaktikai faszerkezetben

Korpusznyelvészet és szövegnyelvészet

A nyelvi változatok vizsgálata

- rétegnyelv, szaknyelv, *genre, register*
- Milyen belső nyelvi jellemzők alapján határozhatók meg?
- Nagy korpuszon vizsgálható igazán
- Hasznos visszacsatolás a korpusznyelvészet számára is

Párhuzamos korpuszok

Forrásszöveg és annak fordítása

- Fordítási megfelelők a mondatok szintjén illesztve
- illesztés a hunalign eszközzel:
  `http://mokk.bme.hu/resources/hunalign`
- Nagy erőkkel folyik kutatás a mondaton belüli egységek illesztésére
- Statisztikai módszerek
- A statisztikai gépi fordítás óriási páruzamos korpuszt igényel
  `http://www.statmt.org/`

Többnyelvű korpuszok

- Európai Parlament
  `http://logos.uio.no/opus/europarl.html`
- *Acquis Communautaire* - EU jogszabálygyűjtemény
  `http://langtech.jrc.it/JRC-Acquis.html`
- Multext-East korpusz és lexikai adatbázis
  `http://nl.ijs.si/ME/V3/`

Magyar-angol korpuszok

- Hunglish korpusz
  `szotar.mokk.bme.hu/hunglish/search/corpus`
- *Acquis Communautaire* - EU jogszabálygyűjtemény
  `http://langtech.jrc.it/JRC-Acquis.html`
- Orwell korpusz
  `http://corpus.nytud.hu/orwell`

Vázlat

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb
  - (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de tetszőleges annotációs rendszer
  - (több azonos is lehet, átfedés is lehet egyes)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb
  - (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb (pl. átfedő elemek)
  - (több azonos szintű elemzés is lehetséges egyszerre)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb
  - (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de tetszőleges annotációs rétegek
  - (több azonos szintű réteg is lehet egyszerre)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb a használata (több szint is lehet, átfedés is lehetséges)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb a használata (több szint is lehet, átfedés is lehetséges)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb a használata (több szint is lehet, átfedés is lehetséges)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb a használata (több szint is lehet, átfedés is lehetséges)

Az annotáció szerepe

Annotáció = a nyelvi elemzés tárhelye

- Az elemzés eredménye az annotációba kerül
- Gyakorlati előny: az online elemzést nem kell mindig ujból futtatni

Két lehetséges megvalósítás

- Az annotáció a tárgynyelvben elhelyezve (*inline*)
  - könnyebb megvalósítani, de korlátozottabb (csak helyes beágyazás lehetséges)
- Az annotáció szövegtől külön szinteken (*stand-off*)
  - nehezebb megvalósítani, de rugalmasabb a használata (több szint is lehet, átfedés is lehetséges)

Egy esettanulmány: előtte

IO[A 4.30-5.30-ig tartó kazetta A oldala]IC
IO[Zene]IC
BO[B Konf]BC *IO[Bemondónő]IC*

Kossuth rádió Budapest, középhullámon 540 *IO[ötszáznegyven]IC*, 1016 *IO[ezerszáztizen-hat]IC*, 1251 ***IO[ezerkettőszázötvenegy]IC*** kilohertzen. Az URH adókon és műholdvevővel rendelkező hallgatóink számára a Hotbird *IO[hotbörd]IC* négyes műholdon, a Duna televízió 7,38 *IO[hét egész harmincnyolc]IC* század megahertzes hangcsatornáján. Felhívjuk figyelmüket, hogy középhullámú és az URH adás szétválásakor a műholdas adókon a középhullámú műsort hallgathatják. 4 *IO[négy]IC* óra 30 *IO[harminc]IC* perc.

Eredeti változat

Egy esettanulmány: utána

```
<div n="3" complete="y" desc="B Konf" type="section">
<add type="info"
value="A 4.30-5.30-ig tartó kazetta A oldala"/>
<event desc="music"/>
<sp>
<speaker>Bemondónő</speaker>
<p>
Kossuth rádió Budapest, középhullámon 540
<add type="pron" value="ötszáznegyven"/>,
1016 <add type="pron" value="ezerszáztizenhat"/>, ...
Felhívjuk figyelmüket, hogy középhullámú és az URH adás
szétválásakor a műholdas adókon a középhullámú műsort
hallgathatják. 4 <add type="pron" value="négy"/> óra
30 <add type="pron" value="harminc"/> perc.
</p></sp></div>
```

XML változat

Vázlat

Lexikai elemzés

Morfológia

- Mára már három teljeskörű rendszer
  - HUMOR – MorphoLogic Kft
  - HUNMORPH – BME MOKK
  - Elekfi-rendszer – MTA NYTI

Lexikai adatbázis

- igei vonzatkeret adatbázis
  - kb. 30 ezer igei keret, felszini esetek, szemantikai jegyek, egyedi lexikai elemek is
- névszói adatbázis
  - év végére kb. 100 ezer névszó nyelvtani, szemantikai jegyekkel kódolva

Lexikai elemzés

Morfológia

- Mára már három teljeskörű rendszer
  - HUMOR – MorphoLogic Kft
  - HUNMORPH – BME MOKK
  - Elekfi-rendszer – MTA NYTI

Lexikai adatbázis

- igei vonzatkeret adatbázis
  - kb. 30 ezer igei keret, felszini esetek, szemantikai jegyek, egyedi lexikai elemek is
- névszói adatbázis
  - év végére kb. 100 ezer névszó nyelvtani, szemantikai jegyekkel kódolva

Szintaktikai elemzés

Elemzők

- Egyelőre részleges eredmények
  - METAMORPHO – MorfoLogik Kft
  - HUNPARS – BME MOKK
  - NP, AP elemző, tagmondat felismerő – MTA NYTI

Szintaktikai adatbázis (*treebank*)

www.inf.u-szeged.hu/projectdirs/hlt/corpus2.htm

- Szeged korpusz
  - 1.200.000 szövegszó hat nyelvi változatból
  - kézzel szerkesztett szintaktikai annotáció

Szintaktikai elemzés

Elemzők

- Egyelőre részleges eredmények
  - METAMORPHO – MorfoLogik Kft
  - HUNPARS – BME MOKK
  - NP, AP elemző, tagmondat felismerő – MTA NYTI

Szintaktikai adatbázis (*treebank*)

www.inf.u-szeged.hu/projectdirs/hlt/corpus2.htm

- Szeged korpusz
  - 1.200.000 szövegszó hat nyelvi változatból
  - kézzel szerkesztett szintaktikai annotáció

Szemantikai elemzés

Wordnet

- Hierarchikus lexikai adatbázis (George Miller, Princeton)
- Mentális lexikon modellje
- EuroWordNet, BalkaNet

Magyar Wordnet

- 40 000 szavas magyar változat
- Jövő év közepére
- Interlingual Index (ILI) – átjárás a többi Wordnet változatokhoz

▸ PWN

Szemantikai elemzés

Wordnet

- Hierarchikus lexikai adatbázis (George Miller, Princeton)
- Mentális lexikon modellje
- EuroWordNet, BalkaNet

Magyar Wordnet

- 40 000 szavas magyar változat
- Jövő év közepére
- Interlingual Index (ILI) – átjárás a többi Wordnet változatokhoz

▸ HWN

Szemantikai annotáció

Névkifejezések annotációja

- Tulajdonnév kifejezések - osztályba sorolva
  - személy-, intézmény-, földrajzi nevek stb.
  - dátum, pénz, mennyiség kifejezések
- szövegekben tömegesen fordulnak elő
- HUNNER projekt (MOKK, Szeged, NYTI)

Vázlat

XML: a dolog veleje

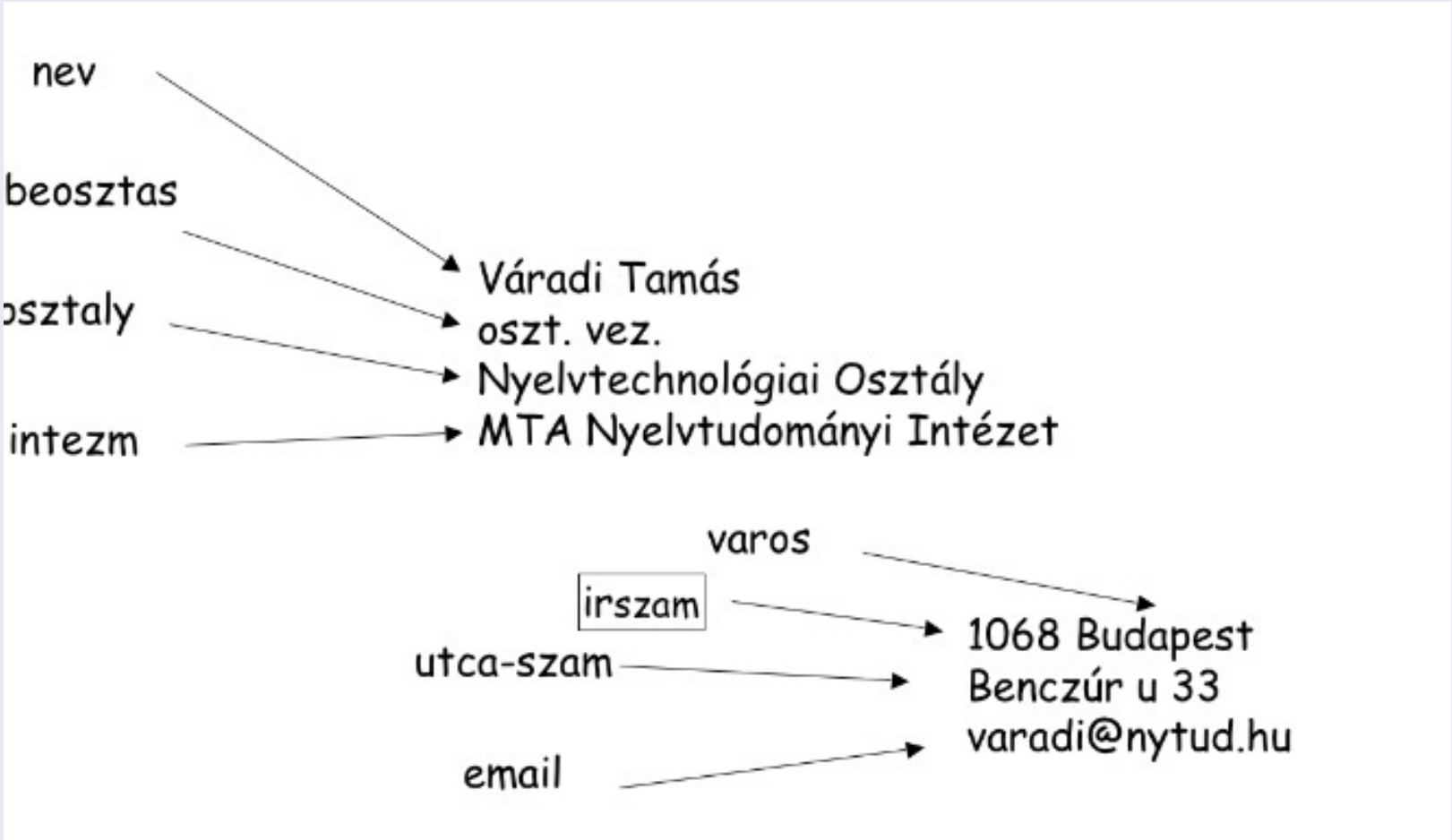
Hagyományos típusú megoldás

Váradi Tamás
oszt. vez.
Nyelvtechnológiai Osztály
MTA Nyelvtudományi Intézet

1068 Budapest
Benczúr u 33
varadi@nytud.hu

XML: a dolog veleje

Hagyományos típusú megoldás



XML: a dolog veleje

XML megoldás

```
        <nev>Váradi Tamás </nev>
  <beosztas>oszt. vez</beosztas>
   <osztaly>Nyelvtechnológiai Osztály</osztaly>
 <intezmeny>MTA Nyelvtudomanyi Intézet</intezmeny>

<irszam>1068</irszam> <varos>Budapest</va
<utca-haz>Benczúr u 33/utca-haz>
<email>varadi@nytud.hu</email>
```

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi feldolgozhatósága lassú

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi feldolgozhatósága lassú

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi feldolgozhatósága lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatbázis-adatcsere szabványa lett

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méreiben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

XML pro és kontra

Mellette

- a formázásban rejtett értelmezést világosan, egyértelműen mutatja
- embernek, gépnek egyaránt jól értelmezhető
- egyszerű eszközökkel kivitelezhető
- csereszabatos adatfájlok

Ellene

- többszörösére növeli a szöveget
- nagy méretben nehezen olvasható
- gépi felhasználása lassú

Konklúzió

- megfelelő szerkesztő programmal jól kezelhető
- adatleírás, adatcsere számára ideális

Vázlat

Korpuszkezelő eszközök

Barátságos, személyi használatra szóló eszközök

- WORDSMITH www.lexically.net/wordsmith
- MONOCONC www.athel.com/mono.html
- PARACONC www.athel.com/para.html

Ipari méretű eszközök

- XAIRA www.oucs.ox.ac.uk/rts/xaira/
- BONITO nlp.fi.muni.cz/projects/bonito/
- IMS CORPUS WORKBENCH

www.ims.uni-stuttgart.de/projekte/CorpusWorkbench

Korpuszkezelő eszközök

Barátságos, személyi használatra szóló eszközök

- WORDSMITH `www.lexically.net/wordsmith`
- MONOCONC `www.athel.com/mono.html`
- PARACONC `www.athel.com/para.html`

Ipari méretű eszközök

- XAIRA `www.oucs.ox.ac.uk/rts/xaira/`
- BONITO `nlp.fi.muni.cz/projects/bonito/`
- IMS CORPUS WORKBENCH

`www.ims.uni-stuttgart.de/projekte/CorpusWorkbench`

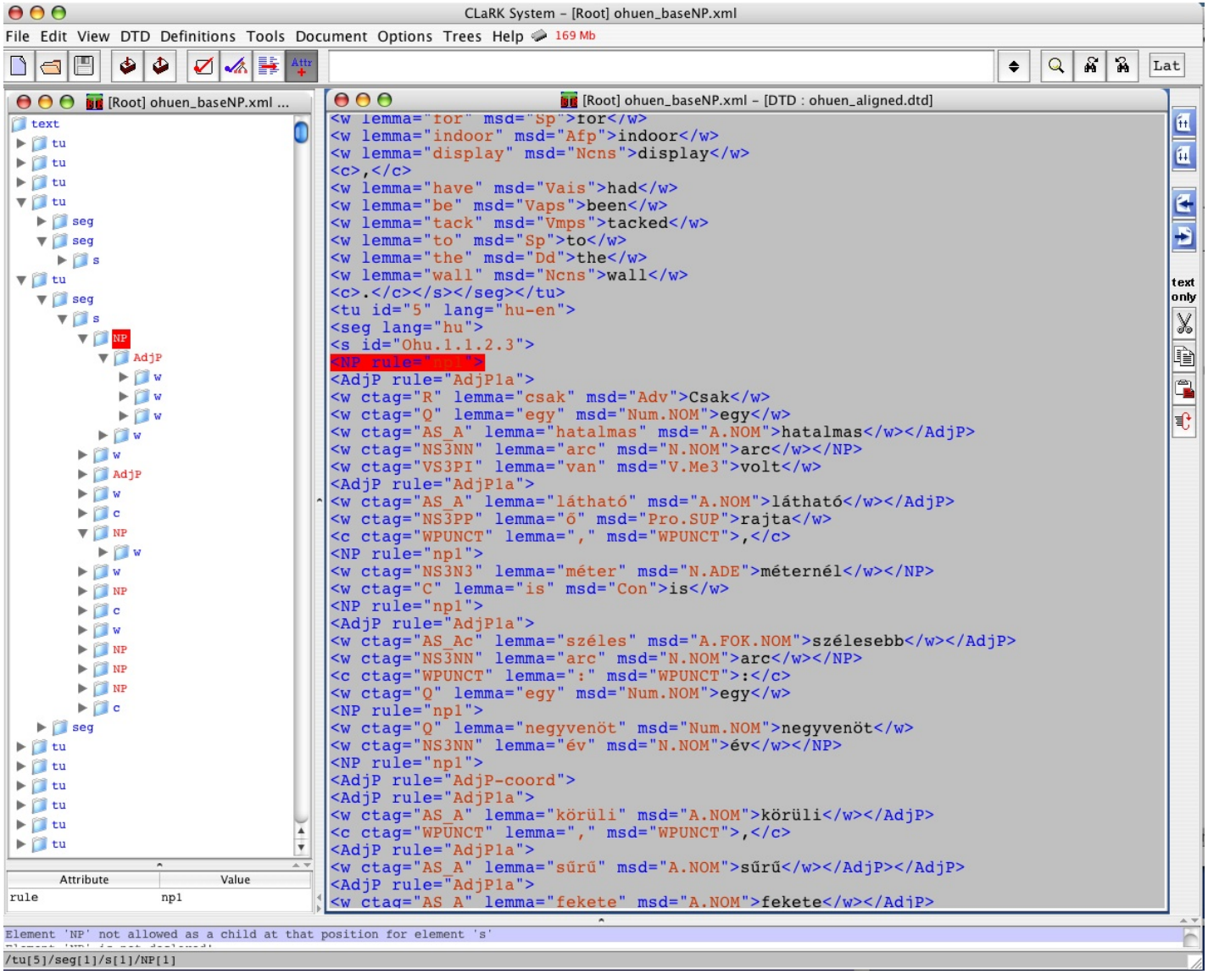
Vázlat

CLaRK - XML szerkesztő és elemző

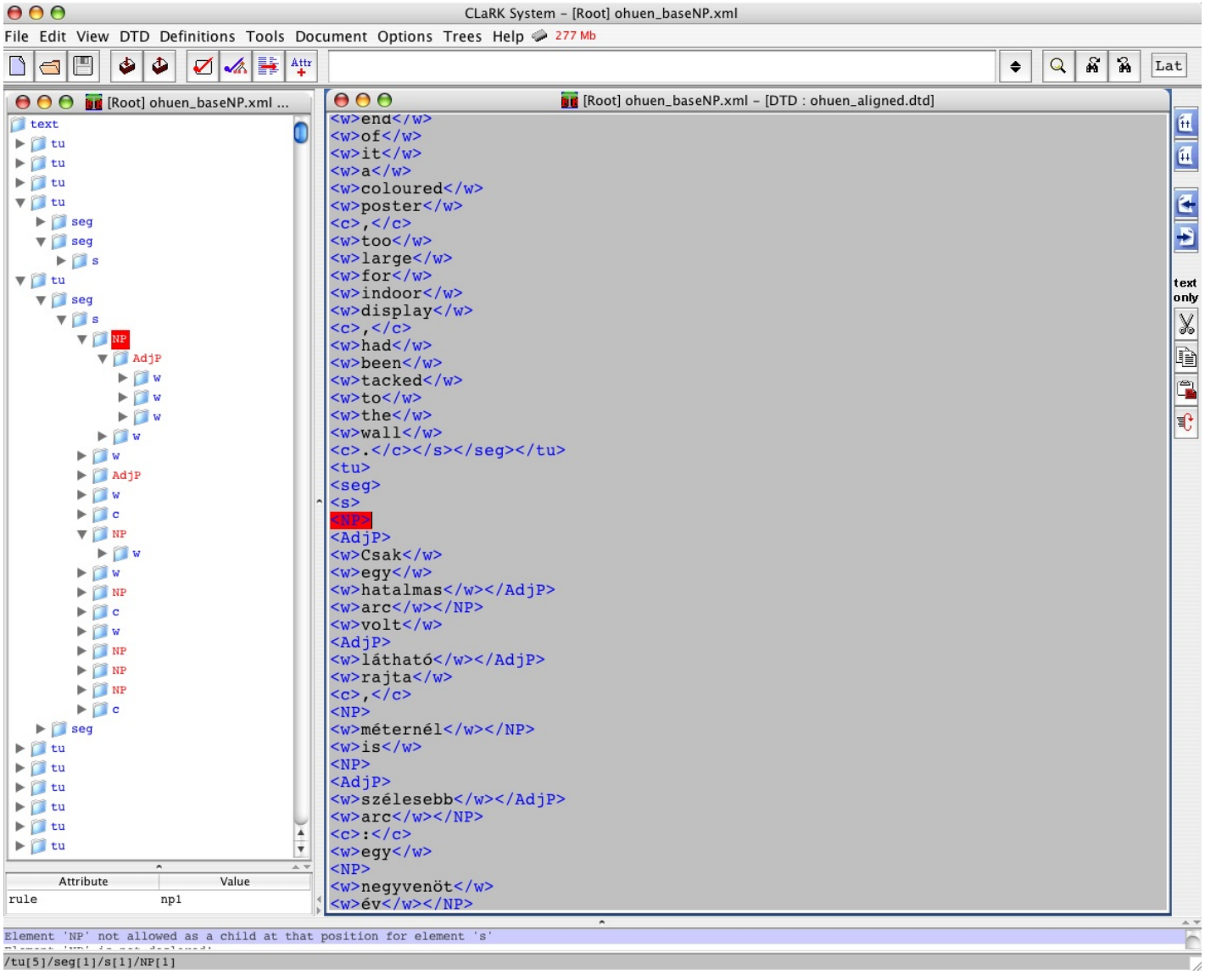
http://www.bultreebank.org/clark/index.html

- ingyenes xml szerkesztő eszköz
- lépcsőzetes, reguláris grammatikák
- véges állapotú technológia
- párhuzamos korpusz kezelésére is alkalmas
- testre szabható, jól kezelhető felület

CLaRK - XML szerkesztő és elemző



CLaRK - XML szerkesztő és elemző



Vázlat

NooJ - végesállapotú keretrendszer

Háttér

- elvek: Maurice Gross, LADL (erős kapcsolat Harris elveivel)
- lokális grammatika – a lokális függőségekre épülő lexikális grammatika
- sok rokonság a konstrukciós grammatikával: lexikon és grammatika egybemosódása, erősen lexikális meghatározottság stb.
- Max Silberztein INTEX majd NOOJ szoftver eszköz

NooJ - végesállapotú keretrendszer

Korpuszkezelő eszköz

- gyors, könnyű kezelés
- felszini alakok és komplex grammatikai részrendszerek egyaránt lekérdezhetők

Grammatika-fejlesztő eszköz

- teljeskörű morfológia
- komoly lexikon, típusba sorolt jegyrendszerrel
- bővíthető, gazdagítható szótári komponens
- lépcsőzetesen futtatható lokális grammatikák
- fejlett grammatikai eszközkészlet
  - lexikai szűrés, jegy egyeztetés, jegy örökítés

NooJ - végesállapotú keretrendszer

Integrált rendszer

- az eszköz nyelvfüggetlen
- akár nulláról felépíthetünk egy grammatikát
- minden egységesen véges állapotú transzducerként működik
- gyors, robusztus
- könnyen kezelhető
  - lexikon és morfológia szövegfájl-ban szerkeszthető
  - nyelvtanok gráfok formájában, intuitív felületen készíthetők
- www.nooj4nlp.net

Magyar változat

Az alap infrastruktúra

- az ÉKSz. szókészletének teljeskörű ragozása
- 80 ezer címszó – kb. 130 m szóalak
- optimalizálás még hátravan
- indulhat a magyar nyelvtanfejlesztő munka!

Érdeklődő partnereket keresünk!

corpus.nytud.hu/NooJ

Magyar változat

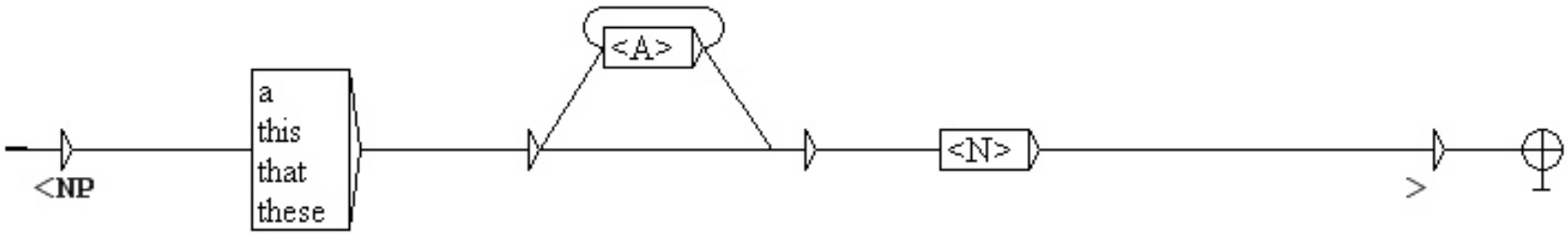
Az alap infrastruktúra

- az ÉKSz. szókészletének teljeskörű ragozása
- 80 ezer címszó – kb. 130 m szóalak
- optimalizálás még hátravan
- indulhat a magyar nyelvtanfejlesztő munka!

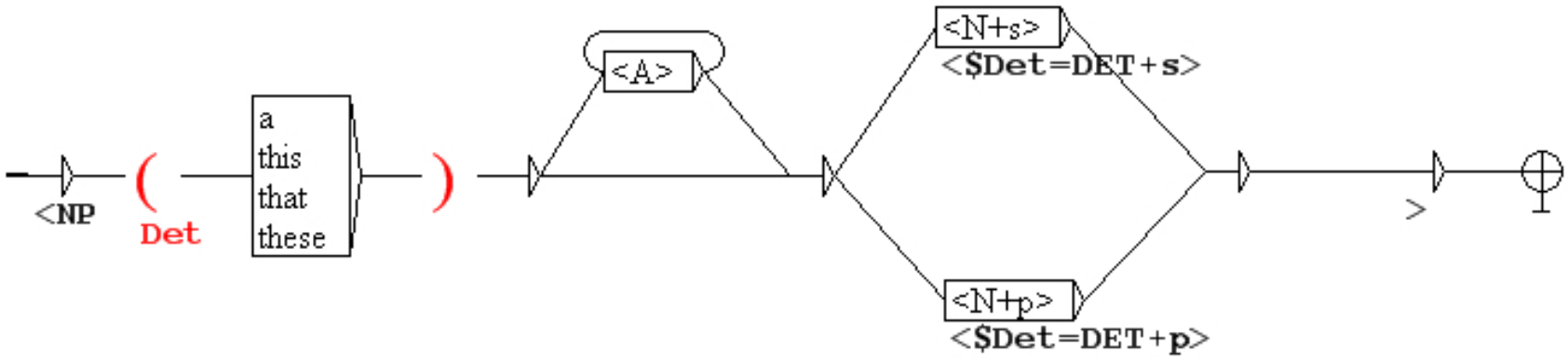
Érdeklődő partnereket keresünk!

corpus.nytud.hu/NooJ

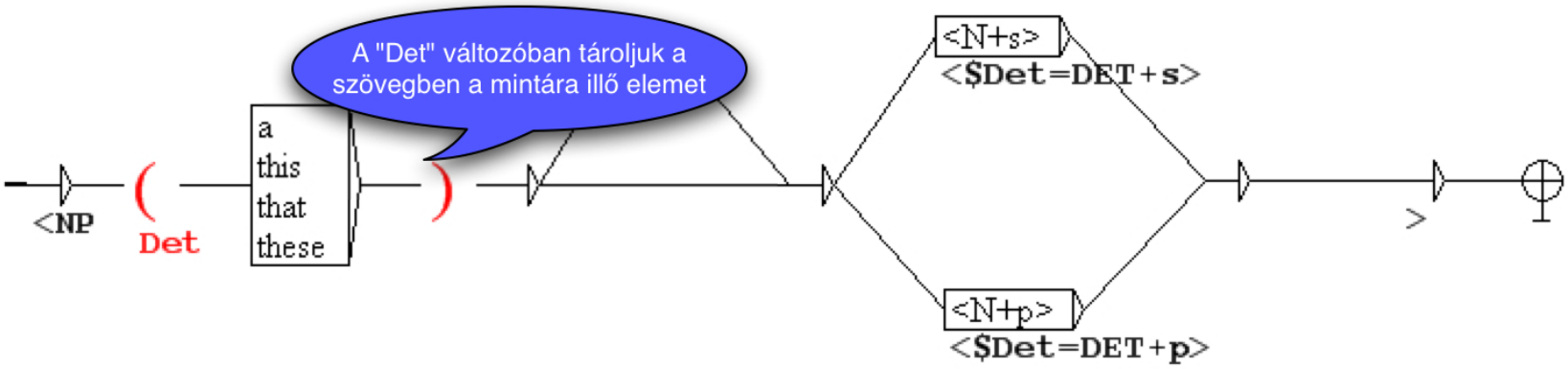
NooJ - lokális grammatika 1,2,3



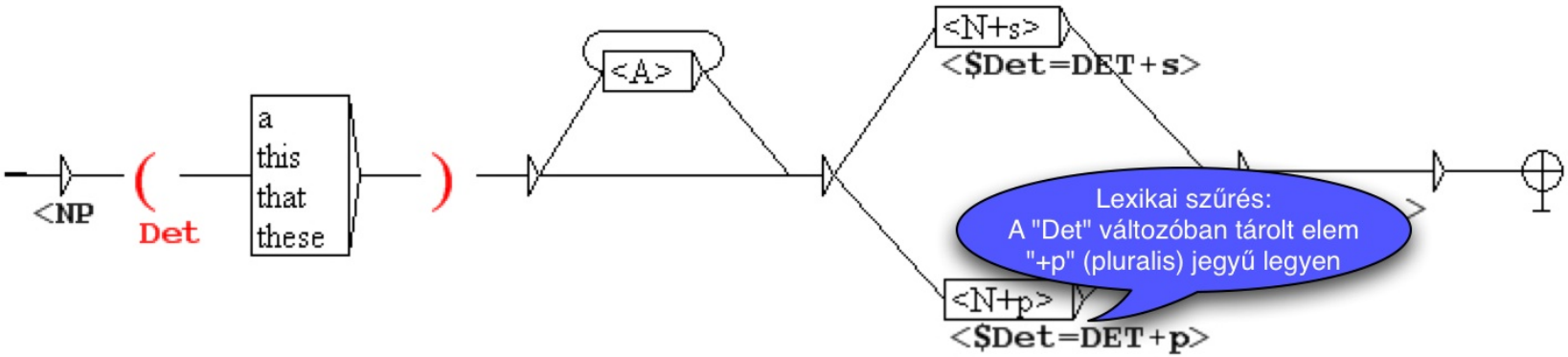
NooJ - lokális grammatika 1,2,3



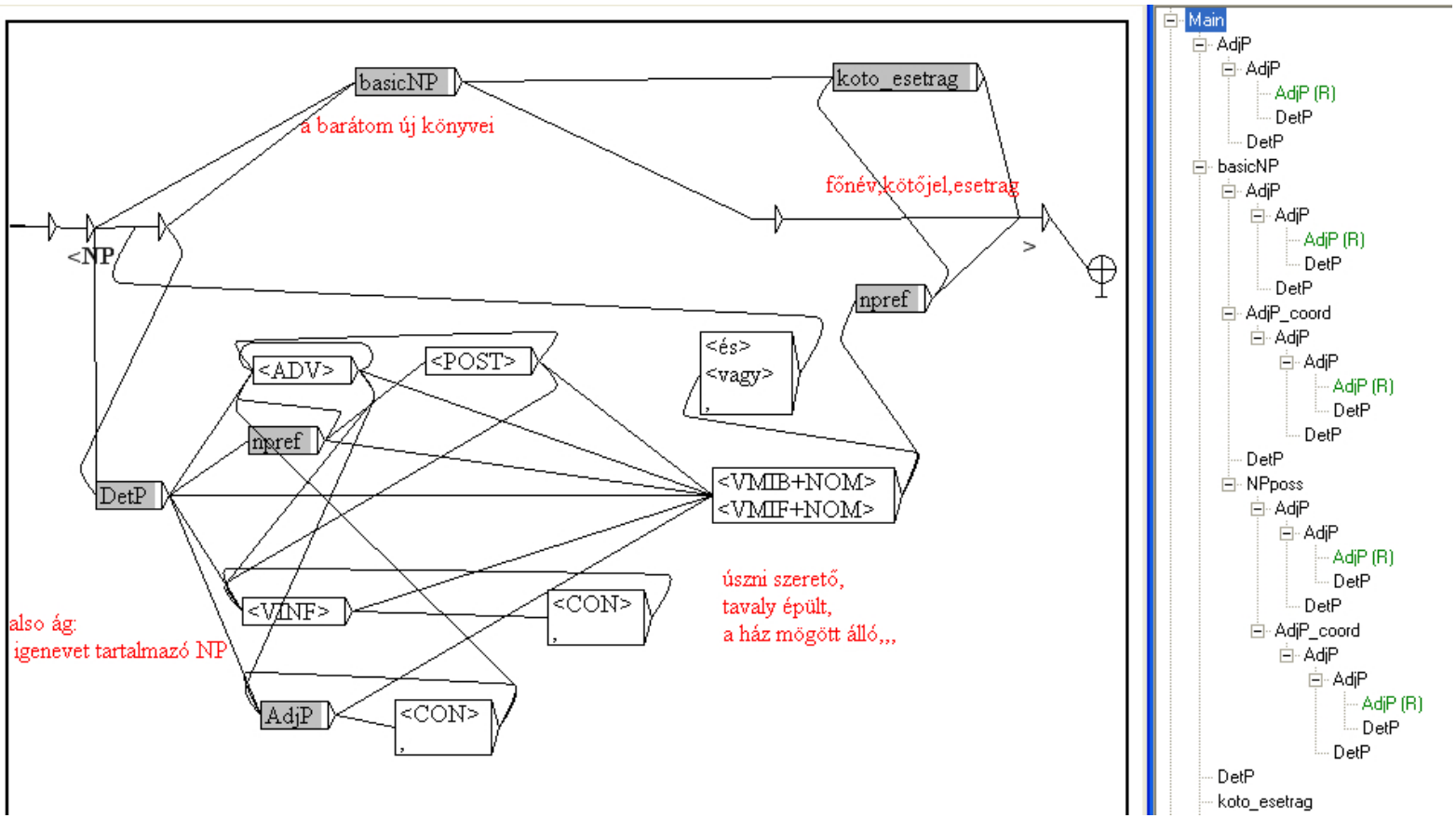
NooJ - lokális grammatika 1,2,3



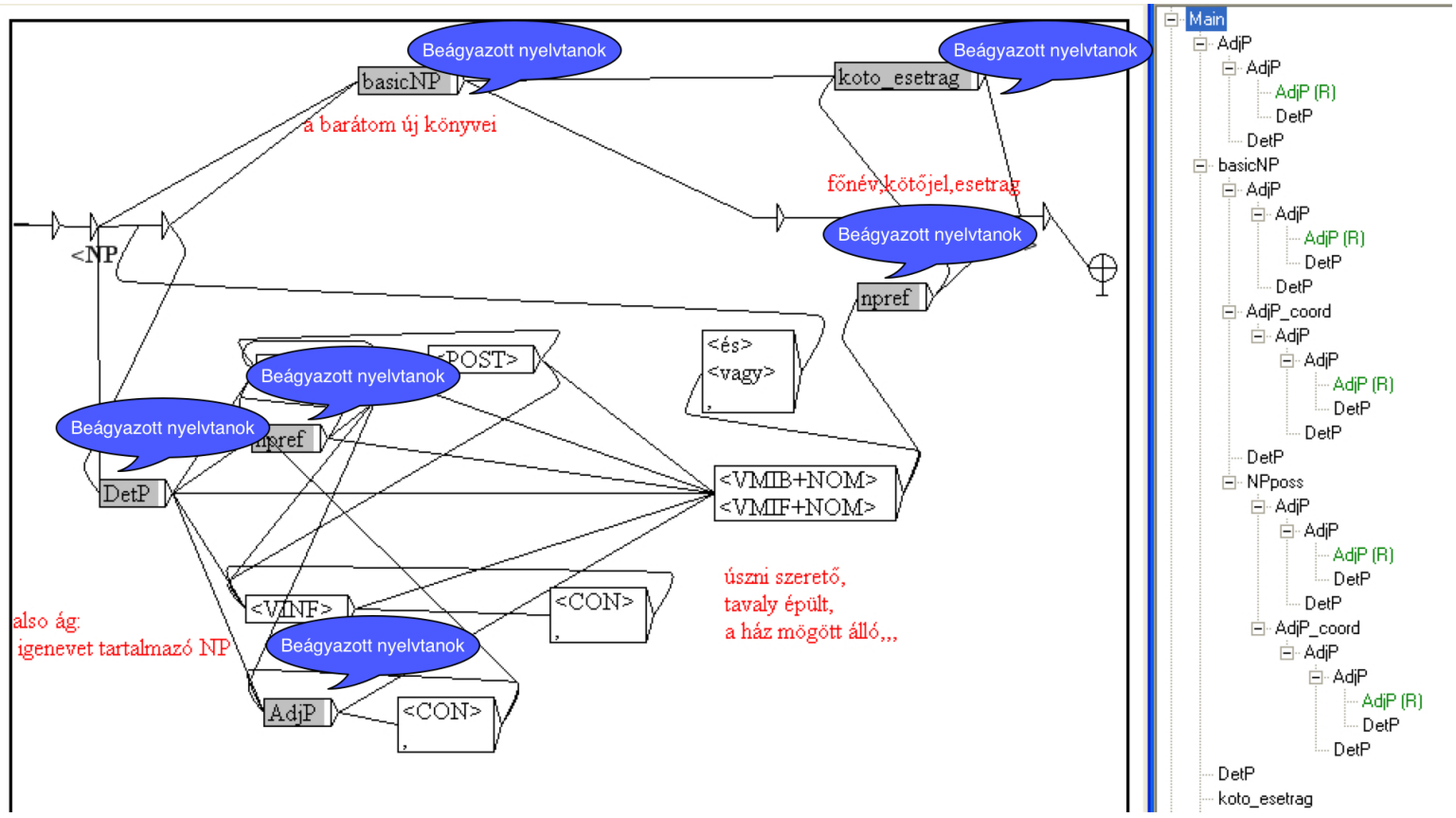
NooJ - lokális grammatika 1,2,3



NooJ - lokális grammatika 1,2,3



NooJ - lokális grammatika 1,2,3



Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni
- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni
- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni

- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni

- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni

- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

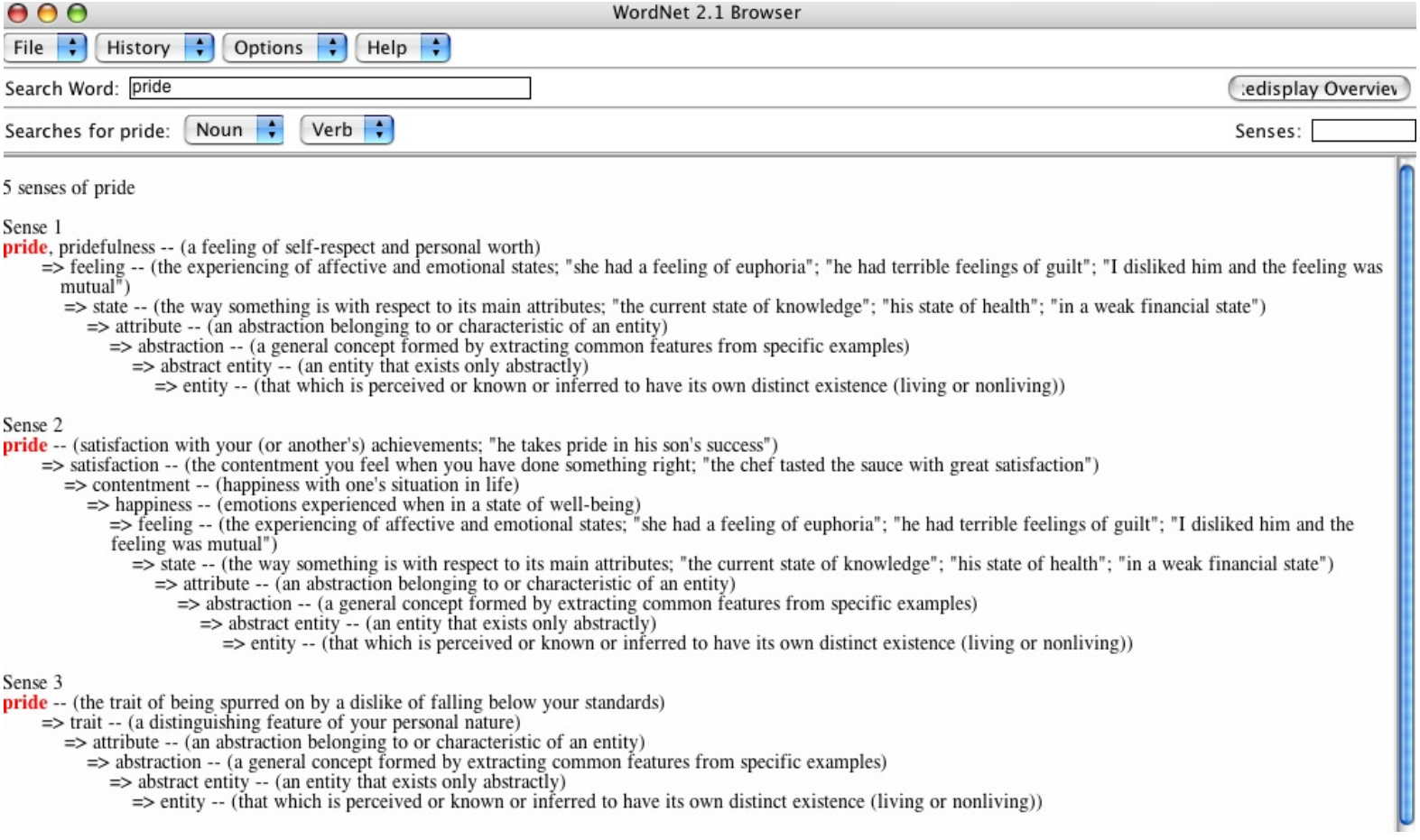
Összegzés

- A nyelvtechnológia elsősorban az informatikai alkalmazások céljait szolgálja
- A végső nyelvészeti kihívás: az emberi szövegértés modellálása
- A korszerű (nyelvi) technológiák alkalmazása a nyelvészeti munkát is képes segíteni

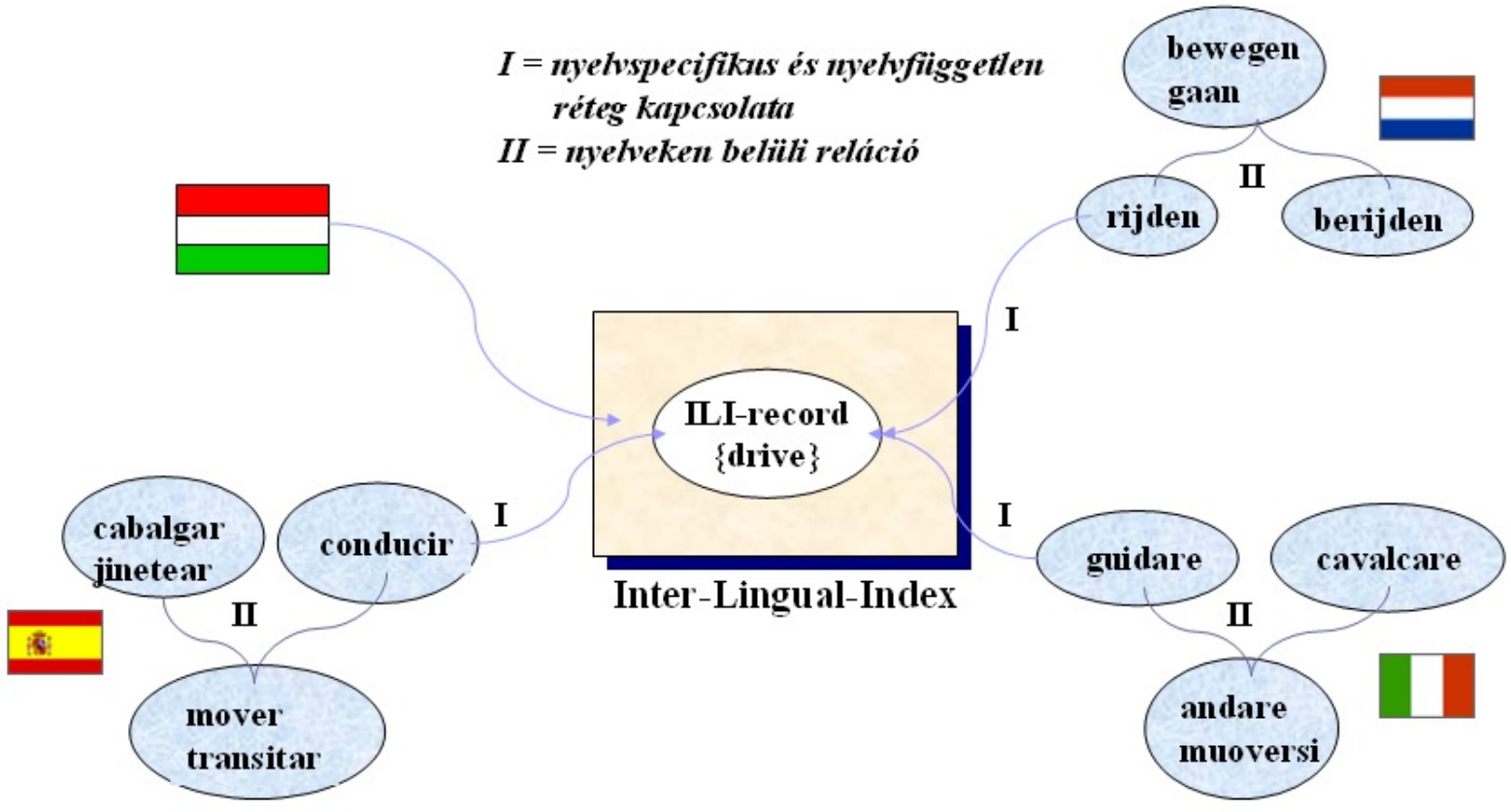
- Alkalmazásának előnyei
  - Pontos, explicit fogalmak és eljárások használatára késztet
  - Azonnali visszajelzés, mérhető eredmények

Köszönöm a figyelmet!

corpus.nytud.hu/people/varadi/talks/nyelvtec.pdf


WordNet 2.1 Browser
File
History
Options
Help
Search Word: pride
Redisplay Overview
Searches for pride: Noun Verb
Senses:
5 senses of pride
Sense 1
pride, pridefulness -- (a feeling of self-respect and personal worth)
=> feeling -- (the experiencing of affective and emotional states; "she had a feeling of euphoria"; "he had terrible feelings of guilt"; "I disliked him and the feeling was mutual")
=> state -- (the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state")
=> attribute -- (an abstraction belonging to or characteristic of an entity)
=> abstraction -- (a general concept formed by extracting common features from specific examples)
=> abstract entity -- (an entity that exists only abstractly)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
Sense 2
pride -- (satisfaction with your (or another's) achievements; "he takes pride in his son's success")
=> satisfaction -- (the contentment you feel when you have done something right; "the chef tasted the sauce with great satisfaction")
=> contentment -- (happiness with one's situation in life)
=> happiness -- (emotions experienced when in a state of well-being)
=> feeling -- (the experiencing of affective and emotional states; "she had a feeling of euphoria"; "he had terrible feelings of guilt"; "I disliked him and the feeling was mutual")
=> state -- (the way something is with respect to its main attributes; "the current state of knowledge"; "his state of health"; "in a weak financial state")
=> attribute -- (an abstraction belonging to or characteristic of an entity)
=> abstraction -- (a general concept formed by extracting common features from specific examples)
=> abstract entity -- (an entity that exists only abstractly)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
Sense 3
pride -- (the trait of being spurred on by a dislike of falling below your standards)
=> trait -- (a distinguishing feature of your personal nature)
=> attribute -- (an abstraction belonging to or characteristic of an entity)
=> abstraction -- (a general concept formed by extracting common features from specific examples)
=> abstract entity -- (an entity that exists only abstractly)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

vissza



P. Vossen nyomán

vissza

Linguistic Data Consortium katalógus

Korpusz típusok

LDC Catalog by Type And Source

| | audio | lexicon | speech | text | video |

Szövegkorpusz típusok

text

| broadcast | broadcast conversation | broadcast news | cellular telephone | dictionary | microphone | news magazine | newsgroups | newswire | parallel | telephone | telephone conversations | transcribed speech | varied | web collection | weblogs |

Mindkettőben uralkodó a *beszédfeldolgozás* igénye

vissza