Magyar Tudományos Akadémia
Nyelvtudományi Intézet

Az annotáció elvei

Oravecz Csaba
MTA Nyelvtudományi Intézet
{oravecz}@nytud.hu

MANYE vitaülés
2006. február 20.

Bevezetés

- Nyelvi erőforrások, szöveges adatbázisok növekvő jelentősége.

Bevezetés

- Nyelvi erőforrások, szöveges adatbázisok növekvő jelentősége.
  - Feladatok: gyűjtés, tárolás, terjesztés, hatékony felhasználás.

Bevezetés

- Nyelvi erőforrások, szöveges adatbázisok növekvő jelentősége.
  - Feladatok: gyűjtés, tárolás, terjesztés, hatékony felhasználás.
  - Eszköz: számítógép.

Bevezetés

- Nyelvi erőforrások, szöveges adatbázisok növekvő jelentősége.
  - Feladatok: gyűjtés, tárolás, terjesztés, hatékony felhasználás.
  - Eszköz: számítógép.
  - Eredmény: nyelvi adatbázisok, korpuszok, elektronikus szótárak stb.

Bevezetés

- Nyelvi erőforrások, szöveges adatbázisok növekvő jelentősége.
  - Feladatok: gyűjtés, tárolás, terjesztés, hatékony felhasználás.
  - Eszköz: számítógép.
  - Eredmény: nyelvi adatbázisok, korpuszok, elektronikus szótárak stb.
- A számítógépes szövegkezelés elterjedése.

Szövegtárolás a számítógépben: a mindennapi módszer

Szövegtárolás a számítógépben: a mindennapi módszer

hegedül

- felépítés: címszó szótári alak;
-
-

Szövegtárolás a számítógépben: a mindennapi módszer

hegedül TN és TS *ige*

- felépítés: címszó szótári alak; bevezető rész;
-
-

Szövegtárolás a számítógépben: a mindennapi módszer

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

- felépítés: címszó szótári alak; bevezető rész; értelmező és szemléltető rész
-
-

Szövegtárolás a számítógépben: a mindennapi módszer

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 megjelenítési konvenciókat, a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
-

Szövegtárolás a számítógépben: a mindennapi módszer

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 megjelenítési konvenciókat, a nyelvet (és a használati útmutatót is) jól ismerő olvasó által
- **procedurális kódolás (markup)**: formázó utasítások, kódok, melyek összekeverednek a dokumentum szövegével

A procedurális kódolás hátrányai

A procedurális kódolás hátrányai

- egy adott megjelenítési formára vonatkozik (pl. nyomtatott oldal)

A procedurális kódolás hátrányai

- egy adott megjelenítési formára vonatkozik (pl. nyomtatott oldal)
- egy adott programcsomaghoz kötődik (pl. . . .)

A procedurális kódolás hátrányai

- egy adott megjelenítési formára vonatkozik (pl. nyomtatott oldal)
- egy adott programcsomaghoz kötődik (pl. . . .)
- a megjelenítési stílus illetve a megjelenítő médium változása a dokumentum teljes átformázásával járhat

A procedurális kódolás hátrányai

- egy adott megjelenítési formára vonatkozik (pl. nyomtatott oldal)
- egy adott programcsomaghoz kötődik (pl. . . .)
- a megjelenítési stílus illetve a megjelenítő médium változása a dokumentum teljes átformázásával járhat
- információ visszakeresése, kinyerése nehézkes

A procedurális kódolás hátrányai

- egy adott megjelenítési formára vonatkozik (pl. nyomtatott oldal)
- egy adott programcsomaghoz kötődik (pl. . . .)
- a megjelenítési stílus illetve a megjelenítő médium változása a dokumentum teljes átformázásával járhat
- információ visszakeresése, kinyerése nehézkes
- nem hordozható formátum

Szövegtárolás a számítógépben: a hatékony módszer

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:
  - a dokumentum szövegének célját rögzíti, s nem a nyomtatásban való megjelenítés módját

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:
  - a dokumentum szövegének célját rögzíti, s nem a nyomtatásban való megjelenítés módját
  - a dokumentum tartalmát elválasztja a megjelenítéstől

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:
  - a dokumentum szövegének célját rögzíti, s nem a nyomtatásban való megjelenítés módját
  - a dokumentum tartalmát elválasztja a megjelenítéstől
  - a dokumentum szerkezetét írja le, és ebben a szerkezetben azonosít egymással meghatározott kapcsolatban álló elemeket $\Rightarrow$ dokumentum típus

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:
  - a dokumentum szövegének célját rögzíti, s nem a nyomtatásban való megjelenítés módját
  - a dokumentum tartalmát elválasztja a megjelenítéstől
  - a dokumentum szerkezetét írja le, és ebben a szerkezetben azonosít egymással meghatározott kapcsolatban álló elemeket $\Rightarrow$ dokumentum típus
  - procedurális markup: "*nyiss egy idézőjelet és válts dőlt betűtípusra*"

Szövegtárolás a számítógépben: a hatékony módszer

- Alapelv: a dokumentum információtartalmának elválasztása a formátumtól.
- A szöveg nem megkülönböztethetetlen bitek és bájtok folyamaként, hanem diszkrét információelemekké darabolva jelenik meg.
- **deskriptív (logikai) markup**:
  - a dokumentum szövegének célját rögzíti, s nem a nyomtatásban való megjelenítés módját
  - a dokumentum tartalmát elválasztja a megjelenítéstől
  - a dokumentum szerkezetét írja le, és ebben a szerkezetben azonosít egymással meghatározott kapcsolatban álló elemeket $\Rightarrow$ dokumentum típus
  - procedurális markup: "*nyiss egy idézőjelet és válts dőlt betűtípusra*"
  - logikai markup: "*a következő dokumentumelem egy példamondat*"

A logikai kódolás előnyei

A logikai kódolás előnyei

- a dokumentum előállítása gyors és hibamentes: a dokumentum mint strukturált objektumok összessége jelenik meg, melyek nem véletlenszerűen jelennek meg, hanem meghatározott kapcsolatban állnak egymással; ellenőrizhető és egyértelmű dokumentumszerkezet

A logikai kódolás előnyei

- a dokumentum előállítása gyors és hibamentes: a dokumentum mint strukturált objektumok összessége jelenik meg, melyek nem véletlenszerűen jelennek meg, hanem meghatározott kapcsolatban állnak egymással; ellenőrizhető és egyértelmű dokumentumszerkezet

- az információ felhasználása, tárolása és más felhasználókkal való megosztása hatékony

A logikai kódolás előnyei

- a dokumentum előállítása gyors és hibamentes: a dokumentum mint strukturált objektumok összessége jelenik meg, melyek nem véletlenszerűen jelennek meg, hanem meghatározott kapcsolatban állnak egymással; ellenőrizhető és egyértelmű dokumentumszerkezet

- az információ felhasználása, tárolása és más felhasználókkal való megosztása hatékony

- a dokumentum hosszú távon is (veszteség nélkül) felhasználható

A logikai kódolás előnyei

- a dokumentum előállítása gyors és hibamentes: a dokumentum mint strukturált objektumok összessége jelenik meg, melyek nem véletlenszerűen jelennek meg, hanem meghatározott kapcsolatban állnak egymással; ellenőrizhető és egyértelmű dokumentumszerkezet
- az információ felhasználása, tárolása és más felhasználókkal való megosztása hatékony
- a dokumentum hosszú távon is (veszteség nélkül) felhasználható
- a dokumentum tartalma számos formátumban rugalmasan megjeleníthető

Szabványos kódolás: XML

Szabványos kódolás: XML

- Extensible Markup Language

Szabványos kódolás: XML

- Extensible Markup Language
- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:

Szabványos kódolás: XML

- Extensible Markup Language
- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:
  - milyen markup használható dokumentumban és hol

Szabványos kódolás: XML

- Extensible Markup Language
- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:
  - milyen markup használható dokumentumban és hol
  - milyen markup kötelező

Szabványos kódolás: XML

- Extensible Markup Language
- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:
  - milyen markup használható dokumentumban és hol
  - milyen markup kötelező
  - hogyan különböztethető meg a markup a szövegtől

Szabványos kódolás: XML

- Extensible Markup Language

- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:

  - milyen markup használható dokumentumban és hol
  - milyen markup kötelező
  - hogyan különböztethető meg a markup a szövegtől
  - mi az alkalmazott markup jelentése.

Szabványos kódolás: XML

- Extensible Markup Language
- **jelölő nyelv** (markup language): a szövegek kódolására használt jelölési előírások egy halmaza. Előírja:
  - milyen markup használható dokumentumban és hol
  - milyen markup kötelező
  - hogyan különböztethető meg a markup a szövegtől
  - mi az alkalmazott markup jelentése.
- Az XML az első háromra vonatkozó szabvány.

A szócikk XML-ben

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>
```

Feldolgozó eszközök

- szerkesztő programok, editorok:
  - <oXygen/> (**http://http://www.oxygenxml.com**)
  - Clark (**http://www.bultreebank.org/clark/**)
  - Xemacs (**http://www.xemacs.org**)
- szerkezetellenőrző programok, validáló elemzők:
  - rxp (**http://www.cogsci.ed.ac.uk/ richard/rxp.html**)
  - SP (**http://www.jclark.com**)
- megjelenítő eszközök, stílusnyelvek (CSS, XSL):
  - **http://www.w3.org/Style/**

További információ

- Általános információ:
  - XML: **http://www.w3.org/XML/**
  - GYIK: **http://xml.silmaril.ie/**
  - Leech, G. (1993): Maxims of Annotation (**http://www.ling.lancs.ac.uk/monkey/↪/ihe/linguistics/corpus2/2maxims.htm**)
- Oktatóanyagok: **http://www.tei-c.org/Tutorials/**
- Programok: **http://www.garshol.priv.no/download/xmltools/**

VÉGE

http://corpus.nytud.hu/manye