A készülő MGTSz adatbázis felépítése

Blaho Sylvia, Sass Bálint & Simon Eszter

MTA Nyelvtudományi Intézet

2010. február 4.

Az előadás vázlata

Outline

A projekt

A projekt:
Magyar **G**eneratív **T**örténeti **Sz**intaxis (MGTSz)

OTKA projekt
É. Kiss Katalin vezetésével
2009.04.01.–2013.03.31.

A projekt célja

elektronikus nyelvtörténeti adatbázis:
a teljes ómagyar és válogatott középmagyar anyag

i. összegyűjtjük és egységesítjük a már meglevő elektronikus nyelvtörténeti anyagokat
ii. a számítógép által olvasható és feldolgozható formára hozzuk az elektronikusan nem elérhetőeket
iii. a betűhű változat mellett előállítunk egy egyszerűsített változatot is
iv. normalizáljuk a szövegeket
v. a korpuszt morfológiailag elemezzük és egyértelműsítjük
vi. a korpusz egy részét szintaktikailag is elemezzük.

A szöveg feldolgozásának szintjei

Az eredeti kódextől a morfológiailag elemzett elektronikusan tárolt szövegig a következő szintek vannak:

(0) fac simile → feldolgozás, kiadás

(1) betűhű átirat → szkennelés, OCR

(2) OCR-ezett szken → javítás, kódolás

(3a) betűhű szöveges elektronikus forma (txt) → egyszerűsítési szabályok

A szöveg feldolgozásának szintjei

(3b) egyszerűsített változat → normalizálás

(4) normalizált alak → automatikus morfológiai elemzés

(5) elemzett forma → automatikus morfológiai egyértelműsítés

(6) egyértelműsített forma

A korpusz felépítése

A korpuszban minden egyes szövegszó mellett szerepelni fognak a következő adatok:

- betűhű forma (3a): *adÿad*
- egyszerűsített alak (3b): *adyad*
- normalizált alak (4): *adjad*
- szótő (6) alapján: *ad*
- morfológiai elemzés (6) alapján: *ad[V.Sub.S2.Def]*

Lekérdezés minden szinten

a lekérdező lényege, hogy *bármely* szinten meg lehet fogalmazni a lekérdezésünket

Példa

„Milyen szavak szerepelnek egy igealak és egy igekötő között?”: (6)

gyakorisági lista a korpusz egy részéből: a szótöveken (6) alapján

az „m̄”-et tartalmazó szavak: (3a)

Kézi kódolás

A kódolók munkájának lényege a (3a) és a (4) alak előállítása.

egységes beviteli formátum: a *kódolási szabályzat* alapján

A kódolási szabályzat

- lókuszjelölők
- Prószéky-kódolás
- mondatrabontás
- a szkriptor javításainak kódolása
- a normalizálás alapelvei

A normalizálás alapelvei 1.

A ma nem létező összes szót, toldalékot, morfológiai konstrukciót meg kell tartani, nem szabad, hogy ilyen információ elvesszen.

(3a)	(4)	értelmezés
villamik	villamik	villámlik/villanik
isa	isa	bizony
iesek	jeszek	jövök

A normalizálás alapelvei 2.

El kell hagyni az összes fonológiai és helyesírási esetlegességet, egységes, amennyire lehet, a mainak megfelelő helyesírásra kell törekedni.

(3a)	(4)
me8den	minden
menden	minden
minden	minden
algyu	ágyú
agyu	ágyú
srumlast	ostromlást

Outline

„Betűhű” változat

- a betűhű változat elkészítésekor nem a szövegek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követtük.
- ezért előfordul, hogy a különböző forrás alapján feldolgozott szövegek más és más egyszerűsítéseket tartalmaznak az eredetihez képest.

„Betűhű” változat

- a betűhű változat elkészítésekor nem a szövegek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követtük.
- ezért előfordul, hogy a különböző forrás alapján feldolgozott szövegek más és más egyszerűsítéseket tartalmaznak az eredetihez képest.

Példa

- *Jókai kódex: ʒ vs. ʒ̓ vs. ʒ́ → a szerkesztő mindhármat ʒ-ként írja át*

„Betűhű” változat

- a betűhű változat elkészítésekor nem a szövegek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követtük.
- ezért előfordul, hogy a különböző forrás alapján feldolgozott szövegek más és más egyszerűsítéseket tartalmaznak az eredetihez képest.

Példa

- *Jókai kódex: ʒ vs. ʒ́ vs. ʒ̋ → a szerkesztő mindhármat ʒ-ként írja át*
- *s vs. ſ, z vs. ʒ, y vs. ẏ vs. ÿ*

„Betűhű” változat

- a betűhű változat elkészítésekor nem a szövegek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követtük.
- ezért előfordul, hogy a különböző forrás alapján feldolgozott szövegek más és más egyszerűsítéseket tartalmaznak az eredetihez képest.

Példa

- *Jókai kódex: ʒ vs. ʒ́ vs. ʒ̋ → a szerkesztő mindhármat ʒ-ként írja át*
- *s vs. ſ, z vs. ʒ, y vs. ẏ vs. ÿ*
- *palatalizált mássalhangzók: ṫ vs. t́*

„Betűhű” változat

- a betűhű változat elkészítésekor nem a szövegek kézzel írott változatát, hanem az általunk használt átirat szerkesztőjének konvencióit követtük.
- ezért előfordul, hogy a különböző forrás alapján feldolgozott szövegek más és más egyszerűsítéseket tartalmaznak az eredetihez képest.

Példa

- *Jókai kódex: ʒ vs. ʒ́ vs. ʒ̋ → a szerkesztő mindhármat ʒ-ként írja át*
- *s vs. ſ, z vs. ʒ, y vs. ẏ vs. ÿ*
- *palatalizált mássalhangzók: ṫ vs. t́*

Egyszerűsített átirat

- célja, hogy
  - az ómagyar szövegek olvasásában nem járatos felhasználó számára megkönnyítse az olvasást,
  - minimalizálja a speciális karakterek használatából adódó problémák előfordulásának esélyét
- elkészítésekor a következő szempontokat tartottuk szem előtt:
  - nyelvészeti relevancia
  - gyakrabban használt karakterek
  - egységesítés

Nyelvészeti relevancia

Az egyszerűsített változatból kimaradnak olyan speciális karakterek és diakritikumok, amelyek paleográfiai, stb. jelentőséggel bírnak, nyelvészetivel azonban nem.

Példa

$$
\begin{aligned}
\int &\rightarrow s \\
ʒ &\rightarrow z \\
\dot{y} &\rightarrow y \\
\ddot{y} &\rightarrow y
\end{aligned}
$$

Gyakrabban használt karakterek

A következő szempont az volt, hogy az egyszerűsített átiratban minél kevesebb, a magyar betűkészletben nem meglévő karakter és mellékjel szerepeljen.

Példa

$\acute{t} \quad \rightarrow ty$

$\acute{o} \quad \rightarrow \ddot{o}$

Gyakrabban használt karakterek

Ahol mégis a magyarban nem haszált betűre volt szükség, előnyben részesítettük az ismertebb, európai nyelvekben gyakran használt karaktereket.

Példa

$\bar{e} \quad \rightarrow \tilde{e}$

Gyakrabban használt karakterek

Ahol mégis a magyarban nem haszált betűre volt szükség, előnyben részesítettük az ismertebb, európai nyelvekben gyakran használt karaktereket.

Példa

$\bar{e} \quad \rightarrow \tilde{e}$

Ez egyrészt az olvasás és a régi magyar példák számítógépes bevitelének megkönnyítését szolgálja (pl. cikkekben, prezentációk segédanyagaiban), másrészt csökkenti a különböző programok fonthasználatából adódó hibák eséyét.

Egységesítés

Azokat a nyelvészetileg releváns karaktereket és diakritikumokat, amelyek különböző nyelvemlékekben eltérő módon jelölik ugyanazt a betűt, illetve tulajdonságot, igyekeztünk egységesíteni.

Példa

$$\begin{aligned} \dot{t} &\rightarrow ty \\ \ddot{t} &\rightarrow ty \\ ty &\rightarrow ty \end{aligned}$$

A legfontosabb szempont az egyszerűsített átirat elkészítésénél azonban az, hogy a betűhű változathoz képest semmilyen, *nyelvészetileg releváns információ* ne vesszen el.

Inkonzisztencia

- egyes kódexek között:

Példa

ġ	*→[g]*	*vs.*	*g*	*→[g]*
g	*→[ɟ]*		*ġ*	*→[ɟ]*

Inkonzisztencia

- egyes kódexek között:

Példa

$\dot{g}$	$\rightarrow[g]$	*vs.*	g	$\rightarrow[g]$
g	$\rightarrow[ɟ]$		$\dot{g}$	$\rightarrow[ɟ]$

- ugyanazon a kódexen belül:

Példa

Müncheni kódex:

tő	g	$\dot{g}$
gyermek	*22*	*0*
hogy	*670*	*0*
míg	*5*	*3*
ország	*137*	*51*

Outline

Korpusz

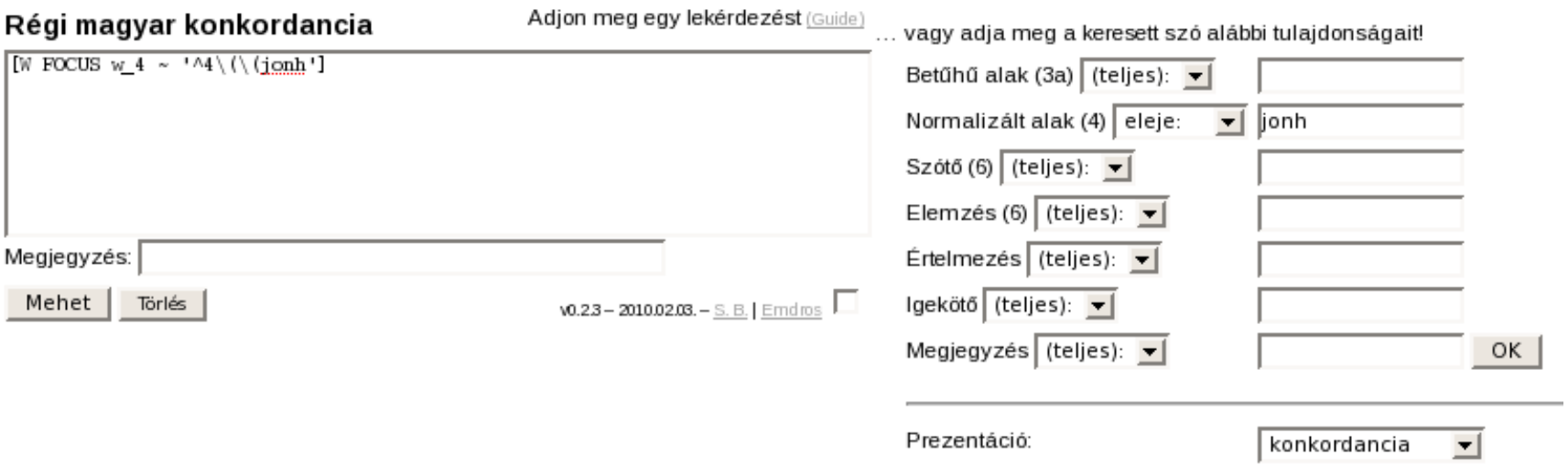
- Formátum: egységes, szabványos, hordozható
  - *XML:* adatbázisszerkezet
    egymásba ágyazott objektumok + attribtumok
    ∼ mondat és szó + a megfelelő szintek adatai szavanként
  - *UTF-8:* különleges karakterek kódolására

- Anyag
  - *Cél:* teljes ómagyar kori anyag
  - *Jelenleg:*

	(3a)	(4)	(6)
kisebb nyelvemlékek	+	+	
Bécsi kódex	+		
Jókai kódex	+		+

Lekérdezőfelület

Eredmény – konkordancia



2010-02-04 02:56:18
Lekérdezés: [W FOCUS w_4 ~ '^4\(\(jonh']
Találati szavak száma: 6 – Futási idő: 3s

[1] MS - 103a5 - 8¢435
eo és menden minden ereſnek erősnek ollian olyan lezen leszen **ionha** **jonha, (szíve)** mit mint pauanak pávának.

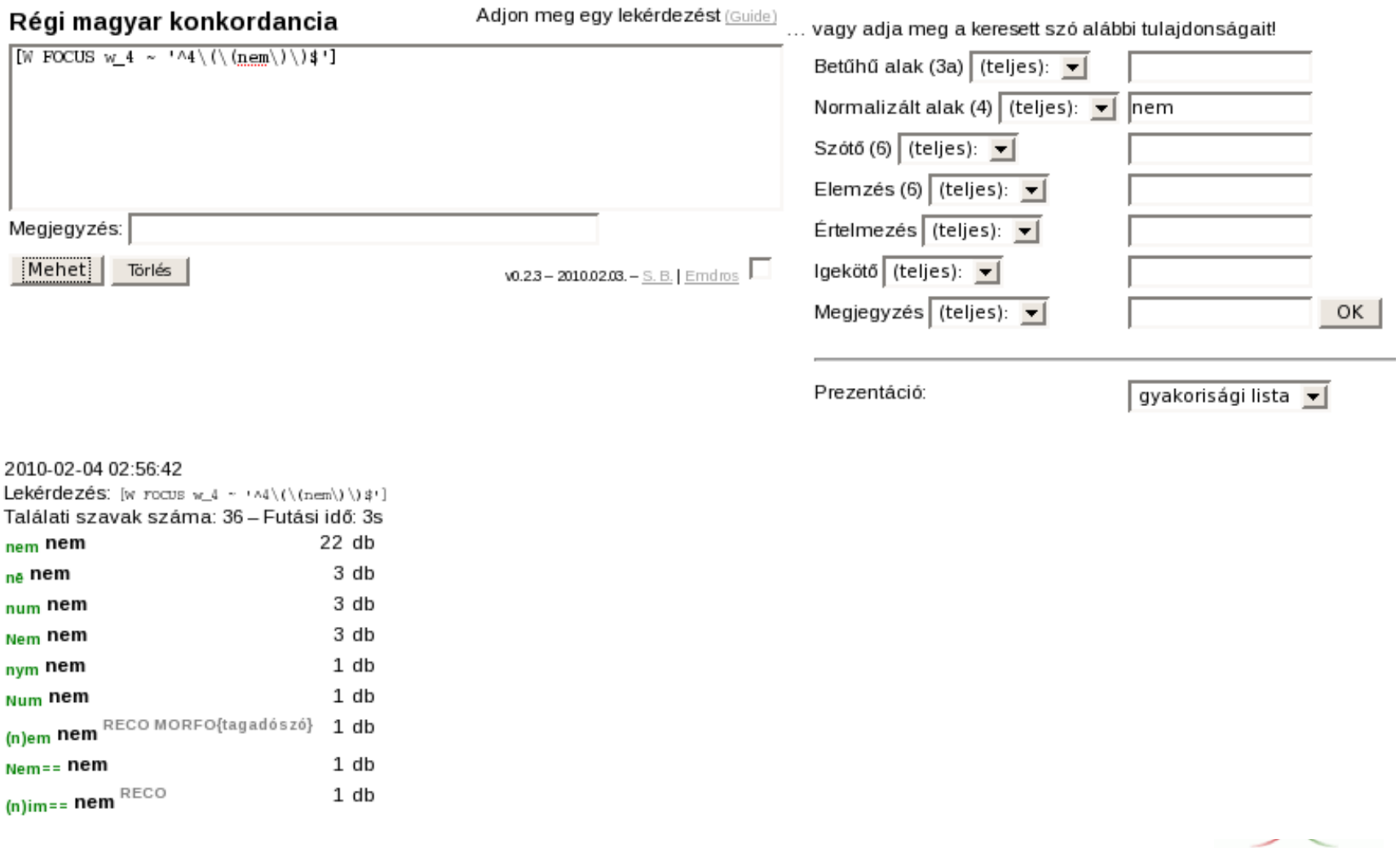
[2] OMS - 5 - 8¢471
en én **iunhum** **jonhom (szívem)** DIFFANA buol búval farad / fárad,

[3] OMS - 10 - 8¢478
en én **iū-hum** **jonhom (szívem)** DIFFANA olelothya alélatja. (alélása) MORFO{noun}

[4] OMS - 18 - 8¢504
en én **iumhumnok** **jonhomnak (szívemnek)** DIFFANA bel bel (belső) bua búja, qui ki (ami) lumha soha nym nem kyul FAIL hyul: hűl.

[5] SzV - 85 - 88285
Nagy nagy bws bús **yonhal** **jonhhal (szívvel)** DIFFANA meg vyssa terenek meg-visszatérének, meg-vissza

Eredmény – gyakorisági lista



Példák (1/4)

- névutók lekérdezése (Hegedűs Vera)

  `[W FOCUS w_6e ~ 'Pp']`

  → 340 db, ebből ragozott 100 db (mind E/3)

- ragozott főnévi igenevek lekérdezése (Tóth Ildikó)

  `[W FOCUS w_6e ~ 'Inf.Px']`

  → 132 db, eloszlás:

	1	2	3
egyes szám	23	23	70
többes szám	5	6	5

Példák (2/4)

„Hol nincs ott a névelő, pedig várnánk?” (Egedi Barbara)

Olyat keresünk, ami *nincs* ott.
→ Adjunk meg konkrét ilyen helyzeteket.

Lekérdezés: definit ige után tárgyesetű főnév

```
[W FOCUS w_6e ~ 'V.*Def']
[W FOCUS w_6e ~ 'N.*Acc']
```

Egy találat

„Es azert ewkewztewk zent ferencz
czudalatost **gÿczerÿuala teremtewtt**”

Példák (3/4)

„Míg a mai magyarban a tagadott igekötős ige (egy-két szerkezetet kivéve) fordított szórendű, a korai ómagyar korban az igekötő az esetek nagyobb részében megelőzi a tagadott igét.” (É. Kiss Katalin)

```
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
[W FOCUS w_6e ~ 'Vpfx']

[W FOCUS w_6e ~ 'Vpfx']
[W FOCUS w_6e ~ 'Mod']
[W FOCUS w_6e ~ 'V\.']
```

Egy találat

„Ver touaba **kÿ nem futott**”

Példák (4/4)

,,Míg a mai magyarban a tagadószó hordozza a tagadást, és a se-névmások csupán a tagadószóval egyeztetett alakok, a korai ómagyar korban a se-névmásoknak is lehetett tagadó erejük.” (É. Kiss Katalin)

Lekérdezés: ’senki/semmi’ után tagadószótól különböző szó

```
[W FOCUS w_6s ~ ’^6s\(\(se[nm][km]i\)\)$’]
[W FOCUS NOT(w_6e ~ ’^6e\(\(Mod\)\)$’)]
```

Egy találat

,,mendenestewlfoguan maganac **semÿtt meg** tarttuan”

Elérhetőség

A lekérdezőfelület szabadon elérhető:

`http://corpus.nytud.hu/rmk`

Köszönjük a figyelmet!