Az Ómagyar Korpusz bemutatása

Simon Eszter

MTA Nyelvtudományi Intézet

2017. január 13.
29. Finnugor Szeminárium

Az előadás vázlata

A projektek

Magyar **G**eneratív **T**örténeti **Sz**intaxis (MGTSz) 1&2

az OTKA támogatásával
(OTKA No. 78074 & OTKA No. 112057)
2009. április – 2013. március & 2015. január – 2018. december

MTA Nyelvtudományi Intézet
projektvezető: É. Kiss Katalin

A projektek célja

MGTSz 1&2

1. elméleti cél: szintaktikai változások rekonstrukciója és vizsgálata
2. számítógépes munka: egy olyan annotált történeti korpusz létrehozása, amely tartalmazza az összes egybefüggő ómagyar szöveget & a korpusz kibővítése középmagyar korból származó Újszövetség-fordításokkal

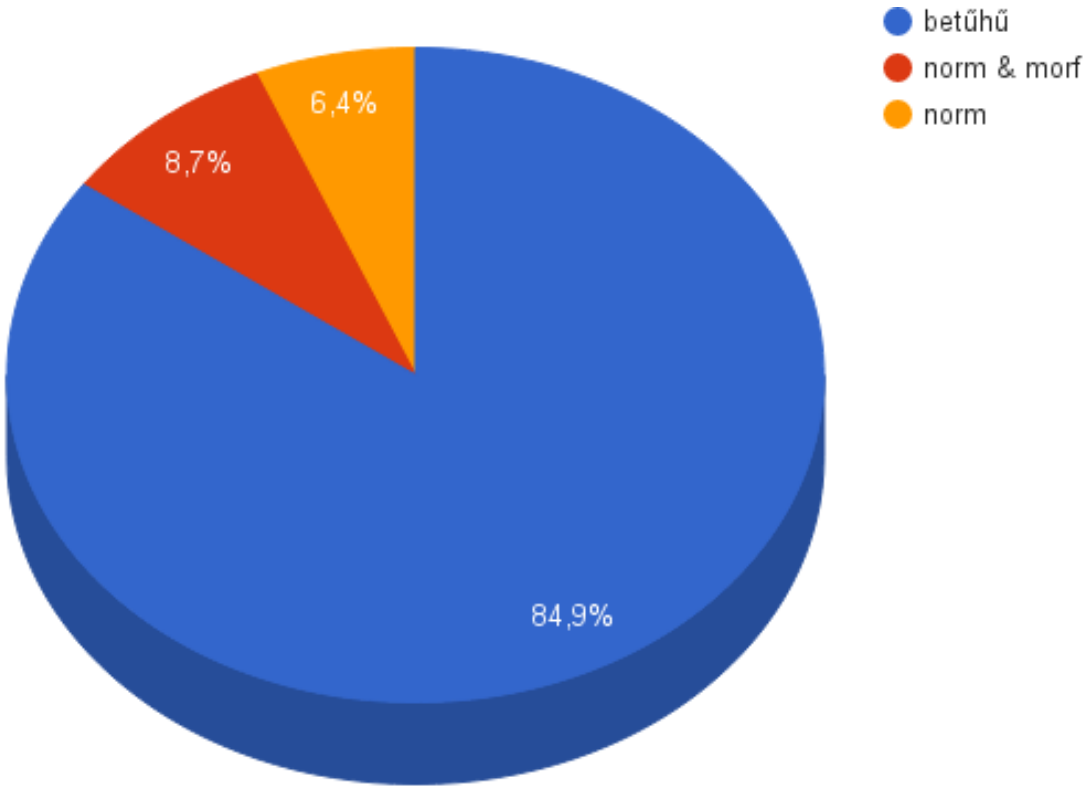
A korpusz anyaga

- 47 ómagyar kódex
- 24 kisebb egybefüggő szövegemlék az ómagyar korból (HB – ME)
- 244 misszilis (Hegedűs–Papp: Középkori leveleink (1541-ig))

- Pesti Gábor (1536)
- Sylvester János (1541)
- Heltai Gáspár (1565)
- Károli Gáspár (1590)
- Káldi György (1626)

A korpusz mérete

2.772.788 token, ebből 419.067 normalizált,
ebből 242.260 morfológiailag elemzett és egyértelműsített



Szövegfeldolgozottsági szintek

(1) kiadott kódex szkennelve
→ OCR

(2) nyers OCR-kimenet
→ *kézi* javítás

(3) betűhű elektronikus forma
→ *kézi* normalizálás

(4) normalizált forma
→ *automatikus* morfológiai elemzés

(5) szótövesített és morfológiailag elemzett forma
→ *félautomatikus* egyértelműsítés

(6) egyértelműsített korpusz

Az annotációs szintek

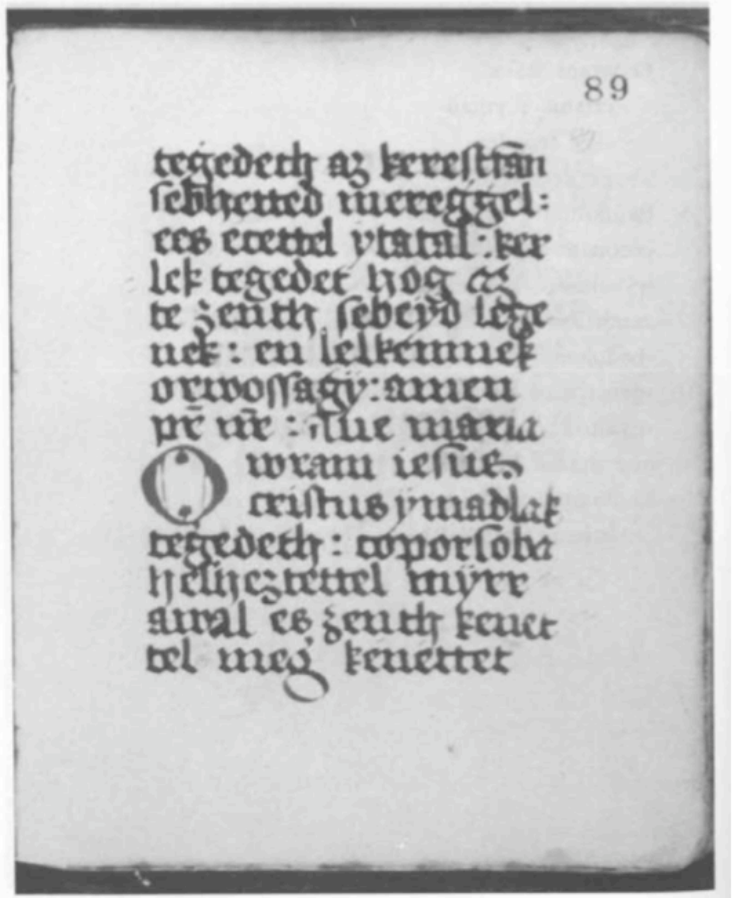
az egyes szövegszavakhoz tartozó annotációs szintek párhuzamosan alakulnak a szövegfeldolgozottsági szintekkel:

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

A korpusz felépítése

oldal	sor	betűhű	norm	lemma	elemzés
001	01	Mÿ	mi	mi	Pro.Nom_gen
001	01	vronknac	Urunknak	Úr	N:P.PxP1.Dat_gen
001	01	iesus	Jézus	Jézus	N:P
001	01	cristusnac	Krisztusnak	Krisztus	N:P.Dat_gen
001	01	gyczeretÿre	dicséretére	dicséret	N.PxS3=i.Sub
001	02	es	és	és	C
001	02	gyczewsegere	dicsőségére	dicsőség	N.PxS3.Sub

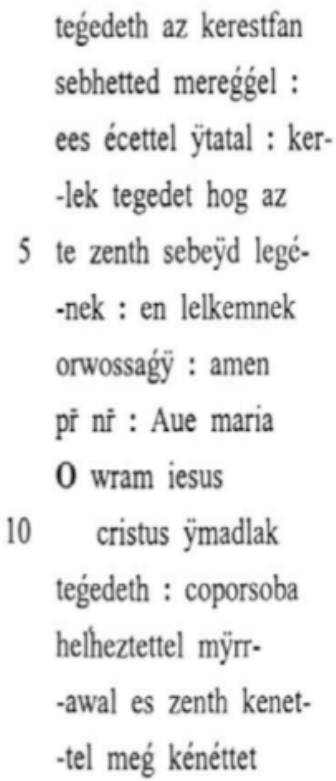
1. Kézzel írott kódexek, nyomtatott kiadások



177
89r

tegedeth az kerestfan
sebhetted meregĝel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pr nr : Aue maria
O wram iesus
cristus ÿmadlak
tegedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meg kénéttet

2. Szkennelés, OCR



tegedeth az kerestfan
sebhetted meregĝel :
ees écettel ÿtatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
př nř : Aue maria
O wram iesus
cristus ÿmadlak
tegedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meg kénéttet

177
89r

tegedeth az kerestfan
sebhetted méreggel :
ees ecettel ytatal : ker-
-lek tégedet hog az
te zenth sebeyd legé-
-nek : en lelkemnek
orwossagy : ámen
pf nf : Aue maria
O wram iesus
eristus ymadlak
tegedeth : coporsoba
hellieztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

3. A betűhű szöveg előállítása

52 latin alapkarakter
42 diakritikus jel
10 szám
15 speciális karakter
34 szövegtagoló és egyéb jel
3 görög betű

Összesen: 156 karakter + ezek kombinációi

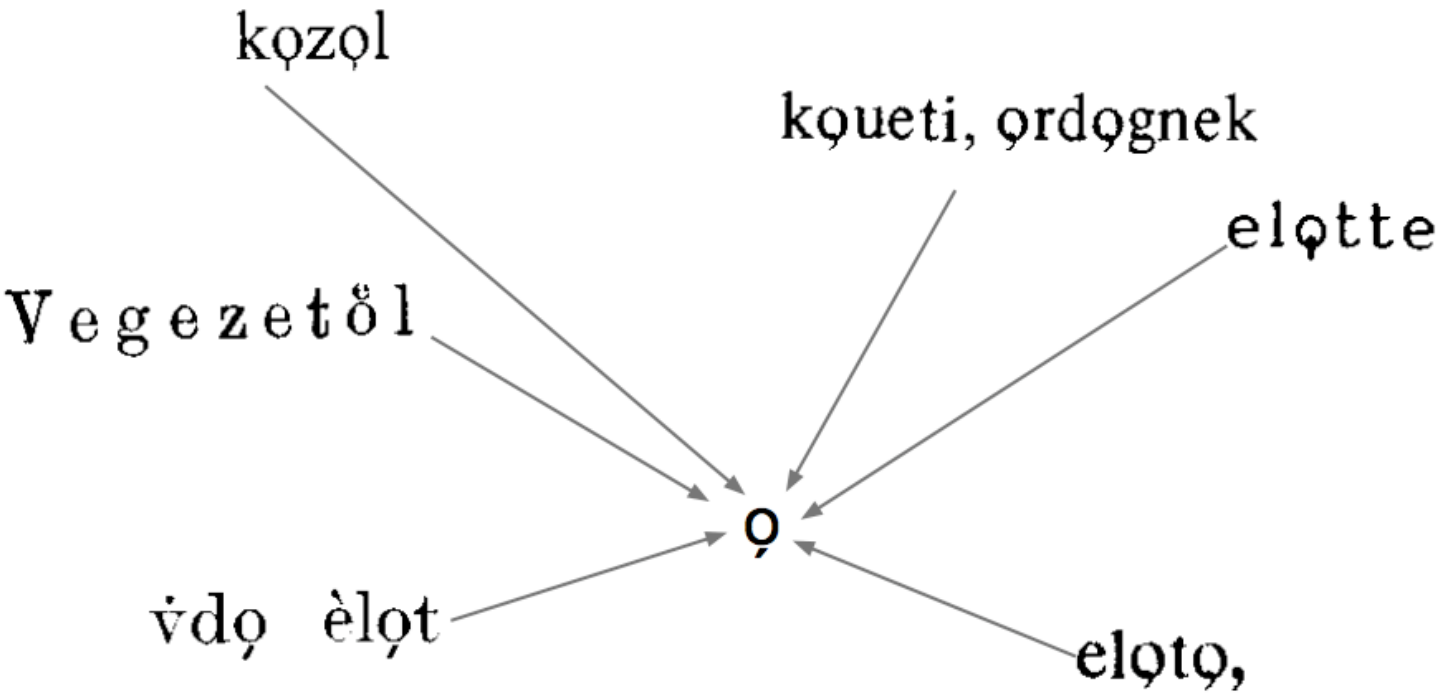
UTF-8 kódolású sztenderd Unicode karakterek

Unicode

- nemzetközi szabvány
- a világ összes nyelvének összes karakterét egy kódolási rendszerbe foglalja
- minden platformon elérhető
- lehetőséget nyújt az alapkarakterek és a diakritikus jelek kombinációjára, pl. $e + ‘ + \bar{} = \check{e}$

az egész korpuszra kiterjedő szigorúan egységes formátum, DE:
van olyan régi karakter, amely a Unicode-ban nincsen reprezentálva
$\longrightarrow$ helyettesítő karakter: $\mathrm{L} \longrightarrow \check{c}$

Hangjelölés-egységesítés



4. Normalizálás

betűhű	normalizált
a varoſba	a városba
ahazi	a házi
annėphėz	a néphez
az arpak	az árpák
ānėp	a nép
a” tew	a tű
a’ nyaar	a nyár
a · mendenhato	a Mindenható

A normalizálás első alapelve

az összes ma már nem létező szót, toldalékot, morfológiai konstrukciót megtartjuk

Példa

ýsa	*pur*	*es*	*chomuv*	*uogmuc*
isa,	*por*	*és*	*hamu*	*vagyunk*
lata	*ǫ*	*napat*	*fèkette*	
látá	*ő*	*napát*	*fekette*	

A normalizálás második alapelve

elhagyjuk az összes fonológiai és helyesírási esetlegességet, vagyis egységes, a mainak megfelelő helyesírásra törekszünk

Példa

meden ∼ *menden* ∼ *mendun* ∼ *mendē* ∼ *mendė* ∼ *miden* ∼ *minden* ∼ *mynden* ∼ *mű́nden* ∼ *mű́ndē* ∼ *mýden* ∼ *mýnden* ∼ *mýndew* ∼ *mÿnden* ∼ *mÿndon* ∼ *mēden* ∼ *mēdèn* ∼ *mēdē* ∼ *mēdėn* ∼ *mėnden* ∼ *mėēden* → *minden*

Tokenizálás és mondatra bontás

Példa

de	*säbädicz==*	*==mk*	*mikët*	*a*	*gonostwl*
de	*szabadíts*	*meg*	*minket*	*a*	*gonosztól*

harmal	*napon*	*halottay bool*	*felthamata*
harmad	*napon*	*halottaiból*	*feltámada*

egmen-@@denic	*ǫ*	*at't'afiat*	*nē*	*zorongat't'a*
egymindenik	*ő*	*atyjafiát*	*nem*	*szorongatja*

Automatikus előnormalizálás

- tokenizálás: egy sor–egy token, a kötőjeles szavak összevonása, az írásjelek leválasztása
- mondatra bontás: minden írásjel után üres sor
- normalizálás: a már normalizált szövegekből készült lista alapján, amely tartalmazza az egy az egyhez megfeleltethető betűhű–normalizált párokat

ezzel a tokenek 63%-át lefedjük, és 49%-nyi időt spórolunk meg a teljesen kézi normalizáláshoz képest

5. Morfológiai elemzés és egyértelműsítés

morfológiai elemzés: a Humor ó- és középmagyarra szabott változatával

egyértelműsítés: kézzel egy webes interfészen keresztül

- a jó elemzés kiválasztható a lehetséges elemzések listájából
- a normalizált szóalak és az elemzés is szerkeszthető, ha szükséges
- az eredmény egy HTML fájl, amit visszakonvertálunk az általunk használt `tsv` formátumra

Univerzális Dependencia és Morfológia

- Humor saját formalizmus → sztenderdizáció → Univerzális Dependencia és Morfológia
- most: a morfológiai elemzés konverziója → később: szintaktikai elemzés hozzáadása
- különbségek az ómagyar és a mai magyar között → különbségek a szófaji és az inflexiós címkékben
- különbségek a két annotációs séma között → bizonyos jelenségeket nem lehetett információvesztés nélkül konvertálni

Metaadatok

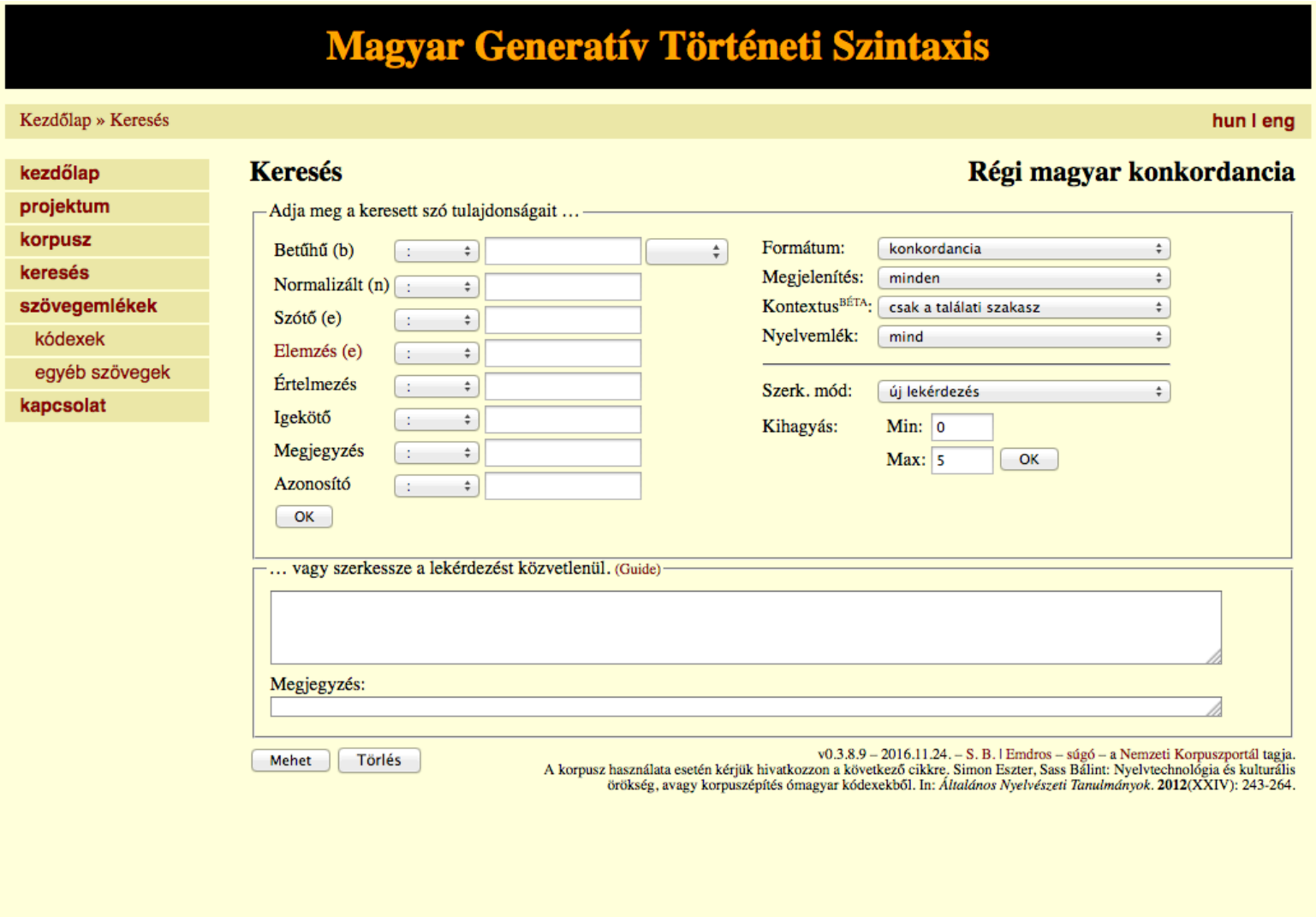
- **lókuszjelölők**
  - az adott szó helye az eredeti kódexben
  - bibliai könyv-, fejezet- és versszámozás
- **értelmezés:** a normalizált alak mai magyarra való “fordítása”
- **megjegyzés:** TITLE, LANG{nyelv}, ADD, RECO, STRIKE, FAIL, FRAG
- **igekötő:** az elvált igekötőt jelöljük az alapige mellett

Lekérdezés minden szinten

- a lekérdező lényege, hogy *bármely szinten* meg lehet fogalmazni a lekérdezésünket
- a lekérdezés eredménye független a lekérdezéstől, vagyis például kereshetünk a normalizált szinten, és megjeleníthetjük a betűhűt
- a lekérdezés eredménye: konkordancia vagy gyakoriségi lista
- a kontextus ablaka állítható
- a szövegemlék specifikálható

http://omagyarkorpusz.nytud.hu/hu-search.html

A felhasználói felület



A lekérdezés eredménye: konkordancia

[384] Konyvecse - 27r - 1/113219

Mÿnden	**fǫldreh**	kÿ meneh	az	ǫ	zǫngesek	,
minden	**földre**	kimene	az	ő	zöngésük	,
minden	**föld**	kimegy	az	ő	zöngés	
N:Pro	**N.Sub**	VPfx.V.Ipf.S3	Det	N:Pro.S3	N.PxP3	

[385] Konyvecse - 27r - 1/113241

zer-@@zed	ǫketh	feÿedelmǫl	mÿnden	**fǫldǫn**	:
szerzéd	őket	fejedelemül	minden	**földön**	:
szerez	ők	fejedelem	minden	**föld**	
V.Ipf.S2.Def	N:Pro.P3.Acc	N.Ess	N:Pro	**N.Sup**	

[386] FestK - 3 - 1/115123

merth	ew	kezeeben	wadnak	**fewldnek**	mýnden	wêgey	:
mert	ő	kezében	vannak	**földnek**	minden	végei	,
mert	ő	kéz	van	**föld**	minden	vég	
C	N:Pro.S3.Nom_gen	N.PxS3.Ine	V.P3	**N.Dat_gen**	N:Pro	N.PxS3.Pl	

A lekérdezés eredménye: gyakorisági lista

Query: `[W FOCUS w_4 ~ 'föld']`
Number of hits: 596 – Elapsed time: 25s

fǫldǫn **földön**	49 db
fǫldnèc **földnek**	44 db
fǫld **föld**	42 db
fǫldèt **földet**	29 db
fǫldebǫl **földjéből**	18 db
fǫldǫn **földön** föld N.Sup	18 db

Párhuzamos Bibliaolvasó

http://parallelbible.nytud.hu/

- bibliai fejezet- vagy vers szinten párhuzamosan lehet megjeleníteni különböző bibliafordításokat
- tartalmaz ó-, közép- és mai magyar bibliafordításokat, plusz a King James Bible-t
- terv: bibliafordítások uráli nyelveken és oroszul is

Köszönöm a figyelmet!

simon.eszter@nytud.mta.hu

http://omagyarkorpusz.nytud.hu/
http://parallelbible.nytud.hu/