Korpuszépítés középmagyar bibliafordításokból – különös tekintettel a Károli-féle Bibliára

Simon Eszter

MTA Nyelvtudományi Intézet

Budapest, 2017. október 18.

Nyelvelmélet és dialektológia 4. – A Károli-Biblia nyelve

Az előadás vázlata

1. A projektekről
2. A korpusz anyaga
3. A korpuszépítés menete
   - A korpuszépítés lépései
   - Nyelvtechnológiai támogatás
4. A korpusz felépítése
5. Korpuszlekérdező eszközök

A projektekről

A projektek

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

az OTKA támogatásával (No. 78074 & No. 112057)

2009. április – 2013. március & 2015. január – 2018. december

MTA Nyelvtudományi Intézet

projektvezető: É. Kiss Katalin

A projektek célkitűzései

Elméleti nyelvészeti:
a magyar nyelv történetét érintő mondattani változások rekonstruálása és leírása

Számítógépes nyelvészeti:

1. a teljes ómagyar anyagot tartalmazó korpusz létrehozása
2. a korpusz kibővítése középmagyar bibliafordításokkal

A korpusz anyaga

MANUSCRIPT &
MIXED MATERIALS

A korpusz anyaga

Az ómagyar anyag

nem foglalkozunk

- szórványemlékekkel
- nyomtatásban még soha ki nem adott szövegekkel

a feldolgozott ómagyar anyag:

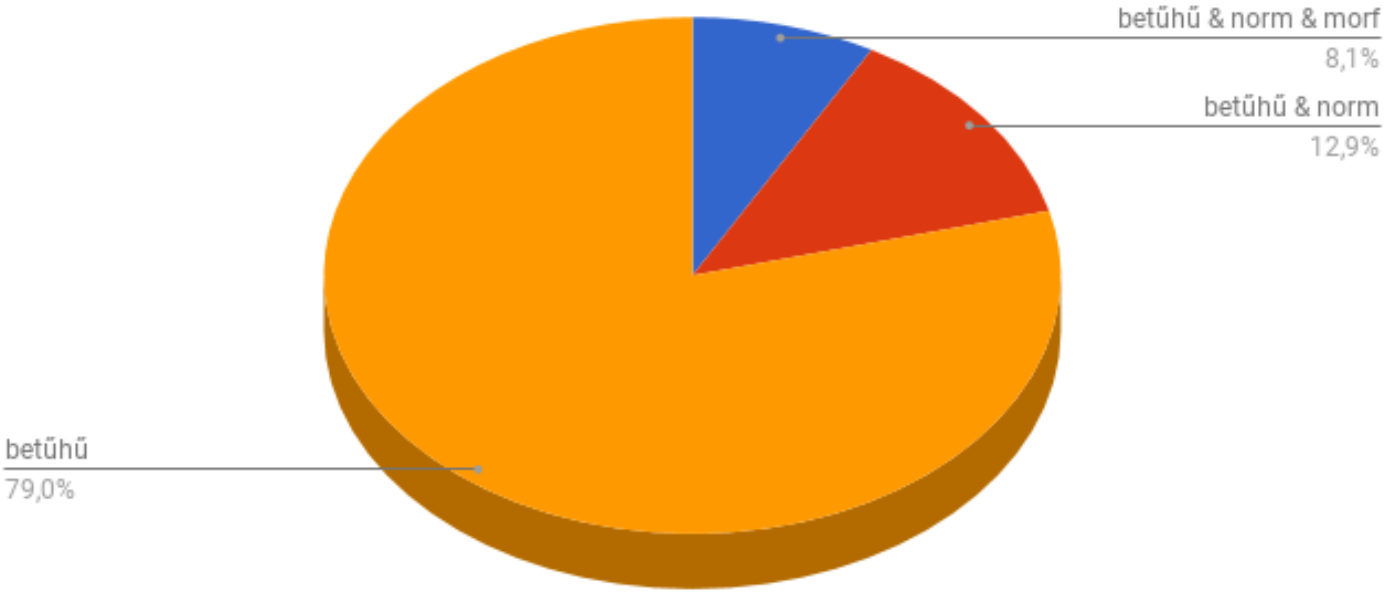
- 47 kódex
- 27 rövidebb szövegemlék
- 244 misszilis (vegyesen az ó- és középmagyar korból)

Középmagyar bibliafordítások

- Pesti Gábor (1536): a 4 evangélium első magyar nyomtatott kiadása
- Sylvester János (1541): az első magyar nyomtatott Újszövetség
- Heltai Gáspár (1565): majdnem az egész Biblia első magyar nyomtatott kiadása
- Károli Gáspár (1590): az egész Biblia első magyar nyomtatott kiadása
- Káldi György (1626): az első magyar katolikus bibliafordítás

A korpusz mérete

3.089.295 token betűhű szöveg (írásjelekkel),
amiből 647.809 token normalizált,
amiből 249.763 token morfológiailag elemzett



Ebből kereshető – betűhű

- az összes ómagyar kódex
- az összes ómagyar rövid szöveg
- az összes misszilis
- az összes középmagyar bibliafordítás (kivéve a Pesti)

Ebből kereshető – normalizált

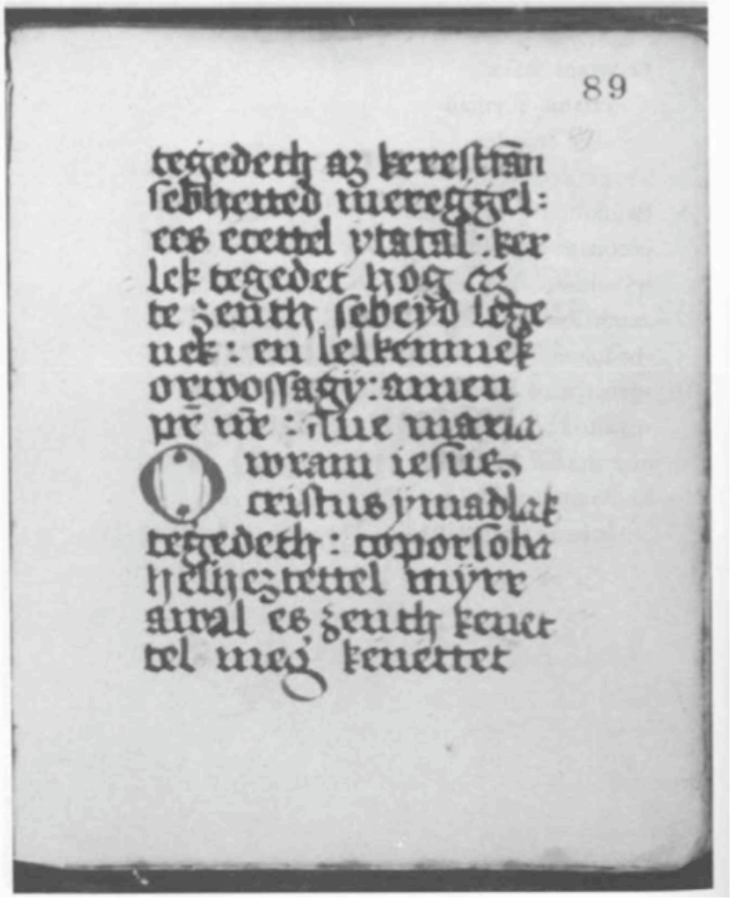
- Bécsi kódex
- Birk-kódex
- Bod-kódex
- Czech-kódex
- Festetics-kódex
- Guary-kódex
- Jókai-kódex
- Jordánszky-kódex (Bevezető + Újszövetség)
- Kazinczy-kódex
- Könyvecse
- Miskolci töredék
- Müncheni kódex
- Székelyudvarhelyi kódex eleje
- az összes rövid szövegemlék (kivéve Csízió)

Ebből kereshető – morfológiailag elemzett

- Festetics-kódex
- Guary-kódex
- Jókai-kódex
- Könyvecse
- Müncheni kódex

A korpuszépítés menete

Kézzel írott kódexek, nyomtatott kiadások



177
89r

tegedeth az kerestfan
sebhetted mereggel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pr̃ nr̃ : Aue maria
O wram iesus
cristus ÿmadlak
tegedeth : coporsoba
helheztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

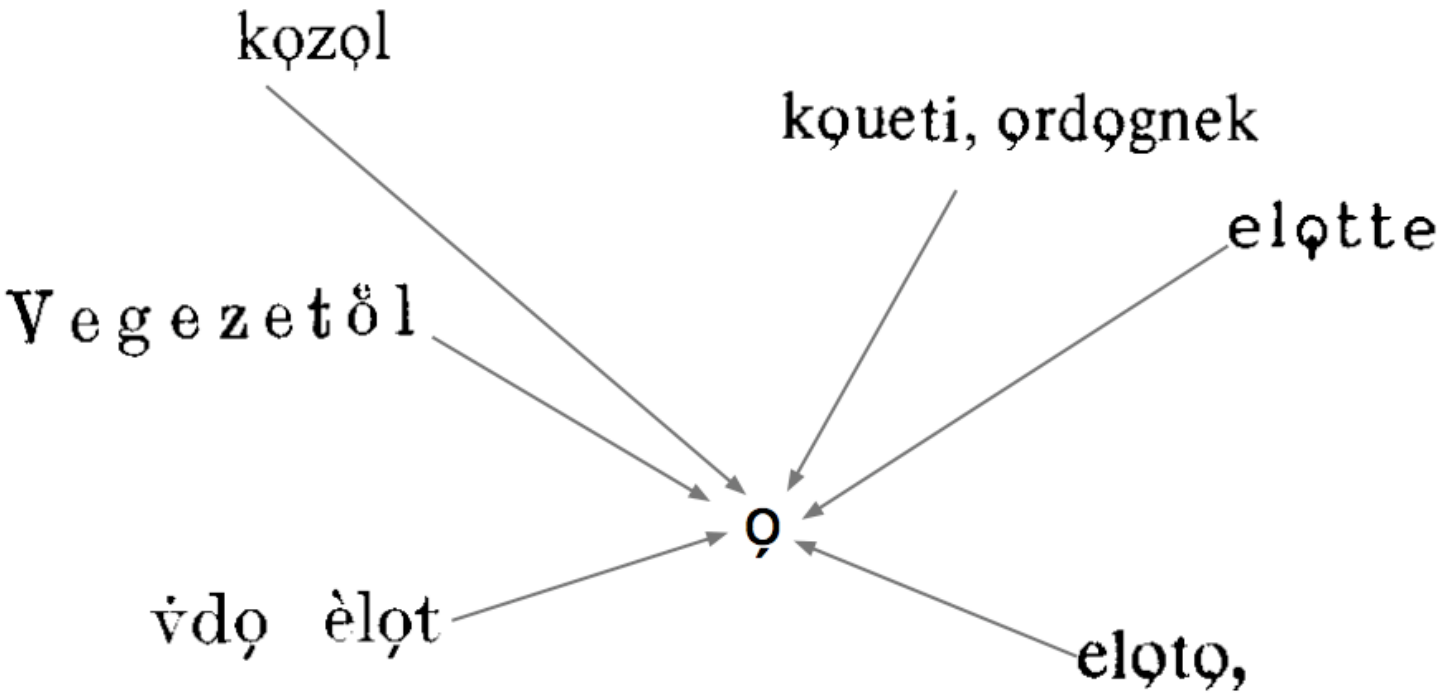
Szkennelés, OCR, javítás

tegédeth az kerestfan
sebhetted meregégel :
ees écettel ÿtatal : ker-
-lek tegedet hog az
te zenth sebeÿd legé-
-nek : en lelkemnek
orwossagÿ : amen
pr̃ nr̃ : Aue maria
O wram iesus
cristus ÿmadlak
tegédeth : 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

Hangjelölés-egységesítés



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	*o*	*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* ∼ *műnden* ∼ *műndē* ∼ *mýden* ∼ *mýnden* ∼ *mýndew* ∼ *mÿnden* ∼ *mÿndon* ∼ *mēden* ∼ *mēdèn* ∼ *mēdē* ∼ *mēdėn* ∼ *mėnden* ∼ *mėēden* → *minden*

Mondatokra bontás

keppen vigaztala meg · Mert va
20 la egÿ frater · zent ferenc zerzety
bevl · hog melÿ frat' jgen nagÿ hÿ
rev[1] vala az ev fratery elevt · es kÿ
ralnak kyralne azzon elevt · Monda
ez frat' ez nagÿ hyrev atÿa · kyral
25 nak · kÿralne azzonnak · es ev zer

Szavakra 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	*o*	*at't'afiat*	*nē*	*zorongat't'a*
egymindenik	*ő*	*atyjafiát*	*nem*	*szorongatja*

Nyelvtechnológiai támogatás

Automatizált lépések:

- a kötőjellel elválasztott szavak összevonása
- az írásjelek leválasztása a szavakról
- tagmondathatár beszúrása minden mondatzáró írásjel után
- átalakítás egy sor – egy szó formátumra
- *előnormalizálás*
- *morfológiai elemzés*

Automatikus előnormalizálás

memóriaalapú: a már elemzett szövegekből készítünk listát, amely a leggyakoribb helyes betűhű–normalizált szópárokat tartalmazza

szabályalapú: környezetfüggő újraíró szabályok

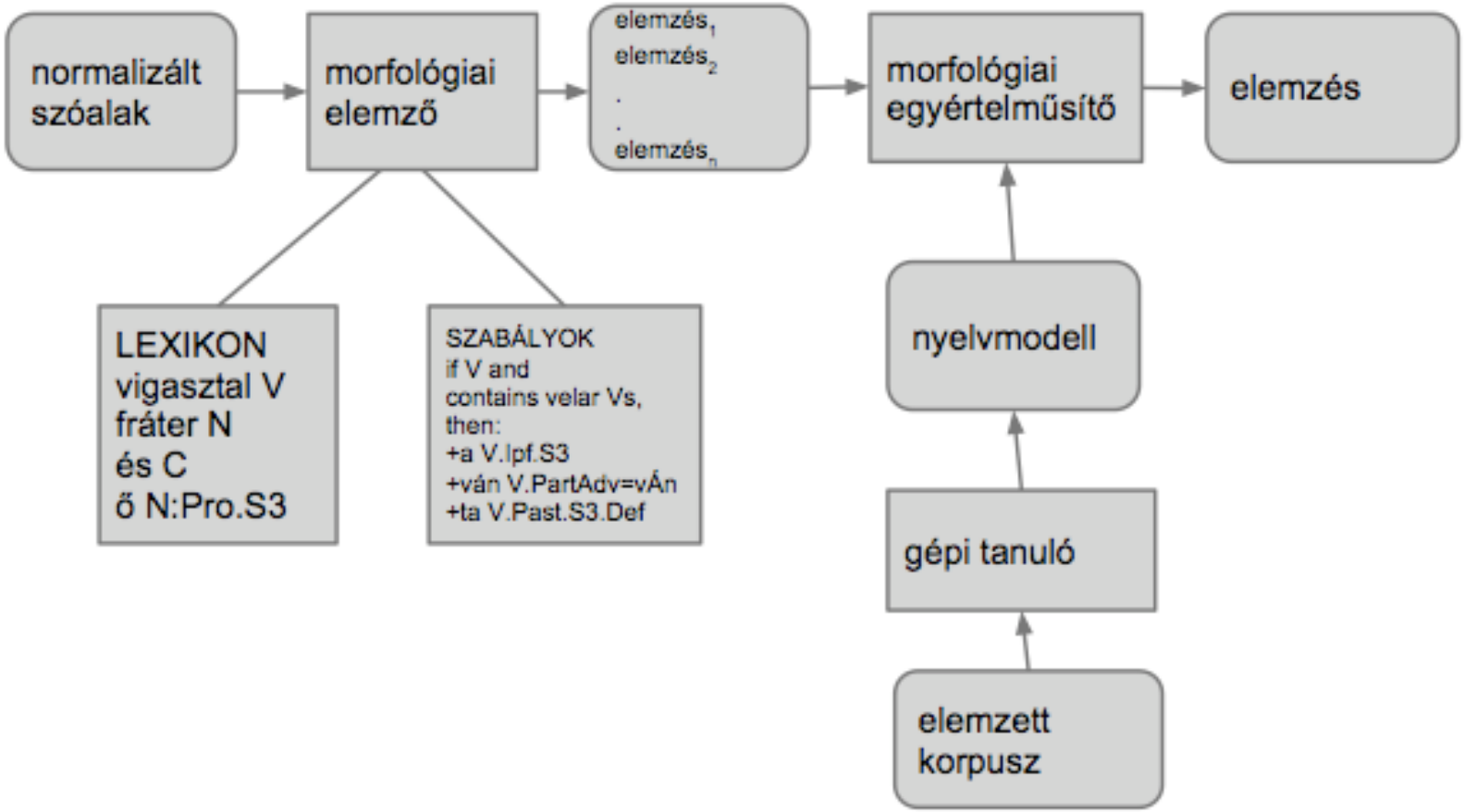
- karakterszintű szabályok: *o$* → *ó*
- tokenszintű szabályok: *le megy* → *lemegy*

Automatikus előnormalizálás – eredmények

könyv	tokenszám (#)	Acc (%)	P (%)	R (%)	F (%)
Máté	19,148	81.04	88.51	70.52	78.49
Márk	12,217	80.60	86.67	66.49	75.25
Lukács	20,810	81.52	85.02	68.07	75.61
János	15,660	83.14	94.74	72.48	82.13
ÖSSZ	67,835	81.59	88.64	69.54	77.94

Acc: normalization accuracy, P: precision, R: recall, F: F-measure

Morfológiai elemzés és egyértelműsítés



Morfológiai elemzés és egyértelműsítés

morfológiai elemzés: **Humor** ('**H**igh speed **U**nification **MOR**phology'): szabályalapú morfológiai elemző, eredetileg a mai magyarra fejlesztve → kibővített lexikonnal és szabályhalmazzal

egyértelműsítés: webes felületen keresztül, ahol az annotátornak egy legördülő menüből kell kiválasztania a helyes elemzést

félautomatikus normalizálás → automatikus morfológiai elemzés → kézi egyértelműsítés

A korpusz felépítése

Szövegfeldolgozottsági szintek

(1) kiadott kódex szkennelve
→ OCR

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

(3) betűhű elektronikus forma
→ *félautomatikus* normalizálás

(4) normalizált forma
→ *automatikus* morfológiai elemzés

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

(6) egyértelműsített korpusz

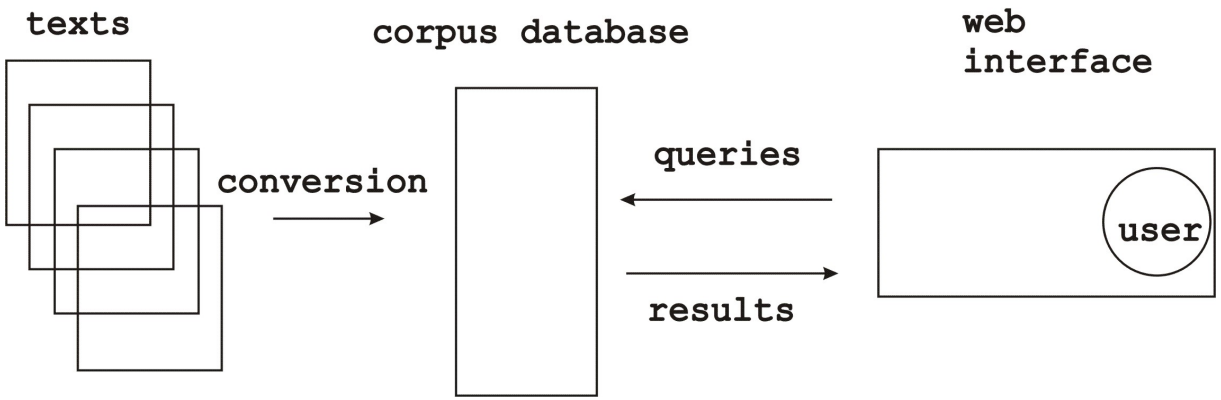
Az annotációs szintek

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

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

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



Korpuszlekérdező eszközök

Korpuszlekérdező eszközök

Régi Magyar Konkordancia:
`http://omagyarkorpusz.nytud.hu/hu-search.html`

Párhuzamos Bibliaolvasó:
`http://parallelbible.nytud.hu/`

Közreműködők

Bácskai-Atkári Júlia
Blahó Sylvia
Egedi Barbara
É. Kiss Katalin
Farkas Judit
Fekete Marcell
Földes Fanni
Gáll Erika
Gerőcs Mátyás
Hegedűs Veronika
Huszthy Bálint
Kacskovics-Reményi Andrea
Kalivoda Ágnes
Kalmár Éva
Kántor Gergely
Mihály Eszter
Mittelholcz Iván
Molnár Roxána
Novák Attila
Oravecz Csaba
Peredy Márta
Pólya Katalin
Sass Bálint
Szeredi Dániel
Szőke Johanna
Tánczos Orsolya
Tóth Ildikó
Varga Dia
Váradi Tamás
Vadász Noémi

Köszönöm a figyelmet!

simon.eszter@nytud.mta.hu

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