Korpuszépítés ómagyar kódexekből

Simon Eszter, Sass Bálint, Mittelholcz Iván

MTA Nyelvtudományi Intézet

2011. december 1.

Az előadás vázlata

1. A projekt bemutatása
2. A korpusz anyagának összegyűjtése
3. Az annotáció kidolgozása
   - Szkennelés
   - OCR
   - A betűhű szöveg
   - Normalizálás
   - Morfológiai elemzés és egyértelműsítés
4. Keresés a korpuszban
5. További feladatok

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

A projekt

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

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

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

elektronikus nyelvtörténeti adatbázis:
a teljes ómagyar (896–1526) és válogatott középmagyar (1526–1772) 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. normalizáljuk a szövegeket
iv. a korpusz egy részét morfológiailag elemezzük és egyértelműsítjük

Korpuszépítési munkálatok

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

- az ómagyar kor több mint 6 évszázadot fog át
- nem volt egységes hangjelölési rendszer, helyesírás
- nagy a szövegek heterogenitása

a korpuszépítés nagyon idő- és munkaigényes folyamat
a sztenderd előfeldolgozó lépések nem végezhetők teljesen automatikusan

Az ómagyar korpusz anyaga

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

a feldolgozandó ómagyar anyag:

- 47 kódex
- 27 rövidebb szövegemlék
- 244 misszilis

összesen kb. 2 millió szövegszó, amelyből több mint 800 ezer már kereshető

Anyaggyűjtés: szöveges formátumok

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

1. különböző források → UTF-8 sima szöveg
2. Számítógépes Nyelvtörténeti Adattár: a kódexbeli szavakat mai magyar változatuk szerint ábécérendbe sorolva adják táblázatos formában → a betűhű, a normalizált és a morfológiailag egyértelműsített szövegszintek rekonstrukciója

Anyaggyűjtés: nem szöveges formátumok

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

```
for codex in codices:
    if codex is short:
        typewriting
    else:
        scanOCRmanualcheck
```

Szövegfeldolgozottsági szintek

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

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

(1)	kiadott kódex szkennelve
	$\longrightarrow$ OCR
(2)	nyers OCR-kimenet
	$\longrightarrow$ *kézi* javítás, kódolás
(3)	betűhű elektronikus forma
	$\longrightarrow$ *félautomatikus* normalizálás
(4)	normalizált forma
	$\longrightarrow$ *automatikus* morfológiai elemzés
(5)	szótövesített és morfológiailag elemzett forma
	$\longrightarrow$ *kézi* egyértelműsítés
(6)	egyértelműsített korpusz

Az annotációs szintek

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

a korpuszban minden egyes szövegszó mellett szerepelnek a következő adatok:

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

Szkennelés

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

néhány kódex beszkennelt verziója megtalálható a MEK-ben, és ezek egy része ún. „szendvics” PDF, DE

- a képek felbontása nem elég jó az OCR-ezéshez
- a mögöttes szöveg nem az ómagyar szövegeken tanított OCR programmal készült, és nincs ellenőrizve

minden kódexet nagy felbontásban beszkenneltünk

OCR

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

kulcsszempont: taníthatóság → Abby FineReader 9.0 Professional edition

az OCR program kiértékelésénél Kniezsa (1952) helyesírási kategorizálását követtük

1. mellékjel nélküli: a latinban nem szereplő magyar hangokat több betű kombinációjával írja le, pl. *cs* → *ch* ∼ *cz* ∼ *chy* ∼ *chi* ∼ *cy*
2. mellékjeles: egy rokonhang betűjének mellékjeles változatával írja le ezeket, pl. *cs* → *č* ∼ *ć*
3. keverék, pl. *cs* → *ch* ∼ *chy* ∼ *cyh* ∼ *c* ∼ *chi* ∼ *č* ∼ *ch'*

Az OCR kiértékelése

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

kódex	helyesírás	token	felismert	WAcc (%)
KulcsK	mellékjel nélküli	36.321	35.258	97,07
MunchK	mellékjeles	74.657	50.790	68,03
CzechK	keverék	11.478	7.910	68,91
–	mai	5.121	5.068	98,97

A betűhű szöveg

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

sztenderd UTF-8 kódolású Unicode karakterek

az egész korpuszra kiterjedő szigorúan egységes formátum, DE: előfordulnak olyan régi karakterek, melyek a Unicode-ban nincsenek reprezentálva → helyettesítő karakter, pl. L → *č*

a szövegeket Prószéky-kódokkal is tároljuk

Heterogén helyesírási rendszer

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

problémás hangjelölés:

- a magyar nyelv hangrendszerének több eleme a latinban ismeretlen, ezek jelölésére új jeleket kellett bevezetni
- nincs egységes helyesírás
- egy kódexet több kéz is jegyezhetett
- egy emléken belül ingadozik egy-egy hang jelölésmódja, pl. *Vylag uilaga* [világ világa]
- kettős hangértéke van egy-egy betűnek, pl. *zerzete zerent* [szerzete szerint]
- néhány betű utalhat magán- és mássalhangzóra is, pl. az *u,v,w* jelölheti az *u,ú,ü,ű,v* hangok bármelyikét

A normalizálás alapelvei 1.

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

A ma nem létező összes szót, toldalékot, morfológiai konstrukciót megtartjuk, vagyis morfémát nem toldunk be, és nem hagyunk el.

betűhű	normalizált	értelmezés
villamik	villamik	villámlik/villanik
isa	isa	bizony
iesek	jeszek	jövök
bèmēuėiec	bemenvéjük	bemenvén T/3.

A normalizálás alapelvei 2.

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

Elhagyjuk az összes fonológiai és helyesírási esetlegességet, vagyis egységes, amennyire lehet, a mainak megfelelő helyesírásra törekszünk. Ez az egységesség elve: egy adott szót mindig ugyanúgy írunk le.

betűhű	normalizált
mēden	minden
menden	minden
minden	minden
algyu	ágyú
agyu	ágyú
strumlast	ostromlást

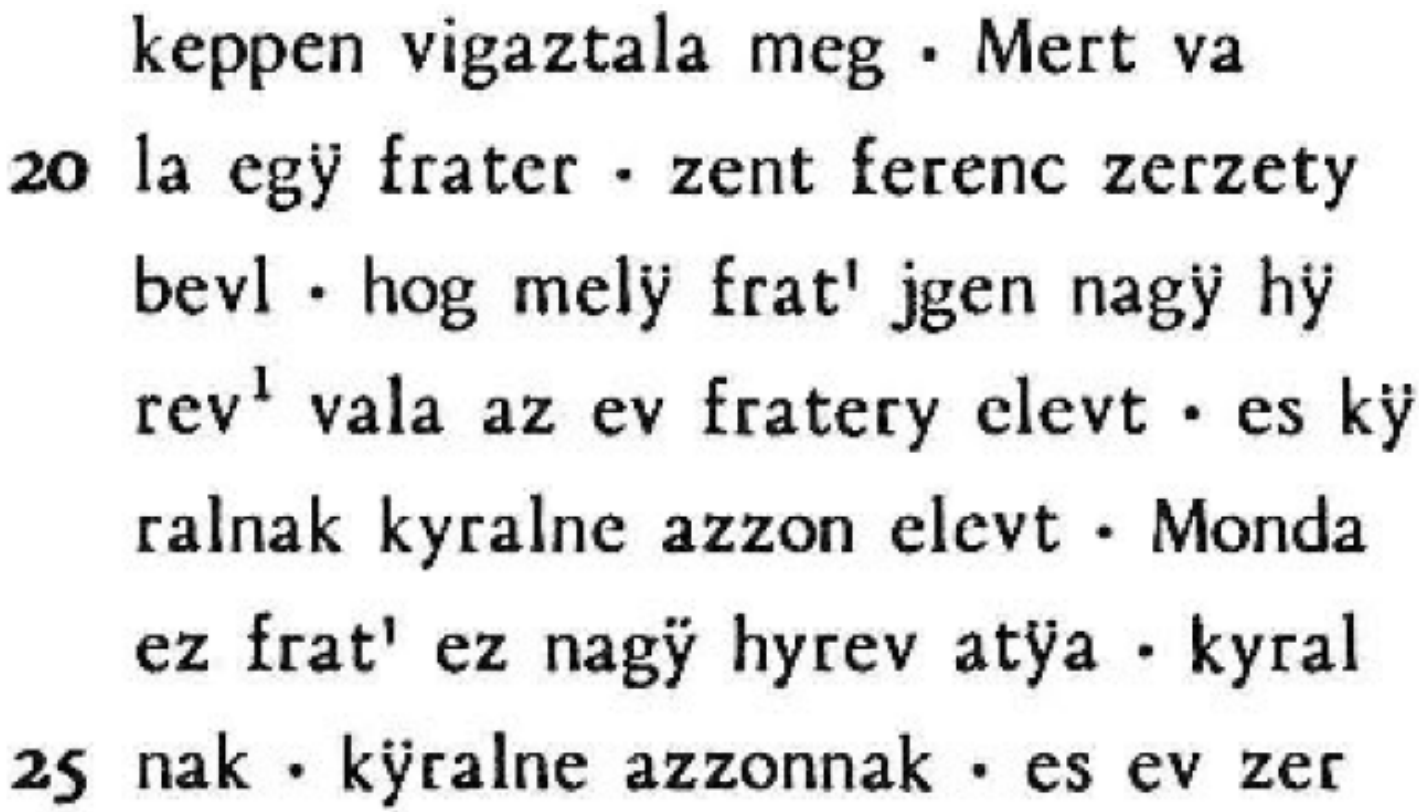
Gépi normalizálás

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

fwl (fül (ear)) →	
-8,80780895229285	föl
-10,7227286786192	fel
-11,0558158154337	fül
-11,2756412387919	fől
honneg (honnét (from where)) →	
-19,1117218113907	honneg*
-19,5230300429664	honnég*
-20,8376176340216	honnét
-21,8538140705439	honyneg*
ygen (igen (yes)) →	
-10,8729908279143	igén
-11,3178857141749	igen
-11,5989613202567	igény
-13,4229320257043	igyen*

Kézi tokenizálás és mondatra bontás

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok


keppen vigaztala meg · Mert va
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
nak · kÿralne azzonnak · es ev zer

Morfológiai elemzés és egyértelműsítés

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

félautomatikus normalizálás $\longrightarrow$ *automatikus morfológiai elemzés* $\longrightarrow$ *kézi egyértelműsítés*

Humor – bővített lexikonnal és szabályhalmazzal

Lekérdezés minden szinten

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

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: (3)

A korpusz anyaga

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

locus	(3)	(4)	ért.	igekötő	megj.
1r	(E)Mbernek	embernek			
1r	elso	első			
1r	ellensege	ellensége			
1r	ez	ez			
1r	velag.	világ,			
1r	ki	ki	ami		DIFFANA
1r	zinetlen	szüntelen			
1r	min-@@ket	minket			
1r	meg ćal.	megcsal,			
1r	es	és			
1r	el hitet,	elhitet.			

A korpuszlekérdező felülete

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

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

A lekérdezés eredménye

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

2011-10-24 14:57:14
Lekérdezés: `[W FOCUS w_4 ~ '^4\(\(jonh']`
Találati szavak száma: 7 – Futási idő: 8s

[1] MS - 103a/5 - 1/130321

eơ	menden	ereſnek	ollian	lezen	**ionha**	mīt	pauanak
és	minden	erősnek	olyan	leszen	**jonha,**	mint	pávának.
					(szíve)		

[2] OMS - 9 - 1/130357

en	**iunhum**	buol	farad /
én	**jonhom**	búval	fárad,
	(szívem)		
	DIFFANA		

[3] OMS - 10 - 1/130364

en	iū-hum	olelothya
én	**jonhom**	alélatja.
	(szívem)	(alélása)
	DIFFANA	MORFO{noun}

További feladatok

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

- a teljes ómagyar anyag betűhű formában való előállítása és kereshetővé tétele
- a különböző korokban kiadott nyomtatott kódexátiratok tipográfiai esetlegességeinek kiküszöbölése
- középmagyar anyagok összegyűjtése és feldolgozása

Közreműködők

A projekt bemutatása
A korpusz anyagának összegyűjtése
Az annotáció kidolgozása
Szkennelés
OCR
A betűhű szöveg
Normalizálás
Morfológiai elemzés és egyértelműsítés
Keresés a korpuszban
További feladatok

Bácskai-Atkári Júlia
Blahó Sylvia
Egedi Barbara
É. Kiss Katalin
Farkas Judit
Gerőcs Mátyás
Hegedűs Veronika
Kacskovics-Reményi Andrea
Kántor Gergely
Mihály Eszter
Mittelholcz Iván
Novák Attila
Oravecz Csaba
Peredy Márta
Pólya Katalin
Sass Bálint
Szeredi Dániel
Szőke Johanna
Tánczos Orsolya
Tóth Ildikó
Váradi Tamás

A projekt bemutatása

A korpusz anyagának összegyűjtése

Az annotáció kidolgozása

Szkennelés

OCR

A betűhű szöveg

Normalizálás

Morfológiai elemzés és egyértelműsítés

Keresés a korpuszban

További feladatok

Köszönöm a figyelmet!