Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

Korpuszépítés ómagyar kódexekből

Simon Eszter

MTA Nyelvtudományi Intézet

2012. április 19.

Az előadás vázlata

1. A projekt bemutatása
2. A korpusz anyagának összegyűjtése
3. A korpusz anyagának feldolgozása
   - Szkennelés
   - OCR
   - A betűhű szöveg előállítása
4. Normalizálás
5. Gépi normalizálás
6. Morfológiai elemzés és egyértelműsítés
7. A korpusz felépítése
8. A korpuszlekérdező eszköz
9. További feladatok

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 (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

Az ómagyar korpusz anyaga

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a feldolgozandó ómagyar anyag:

- 48 kódex
- 27 rövidebb szövegemlék
- 244 misszilis

összesen kb. 2 millió szövegszó, amelyből több mint 850 ezer már kereshető

Anyaggyűjtés: szöveges formátumok

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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
   - sorbarendezés a lelőhely alapján
   - átalakítás UTF-8 kódolású tab separated sima szöveggé
   - a normalizált szóalakok rekonstrukciója
   - a kódok átalakítása az általunk használt morfológiai elemző kimeneti formalizmusára
   - bizonyos nyelvi jelenségek átkódolása

Számítógépes Nyelvtörténeti Adattár

Korpuszépítés ómagyar kódexekből

Simon Eszter

Bemutatás

Anyaggyűjtés

Feldolgozás

Szkennelés

OCR

Betűhű szöveg

Normalizálás

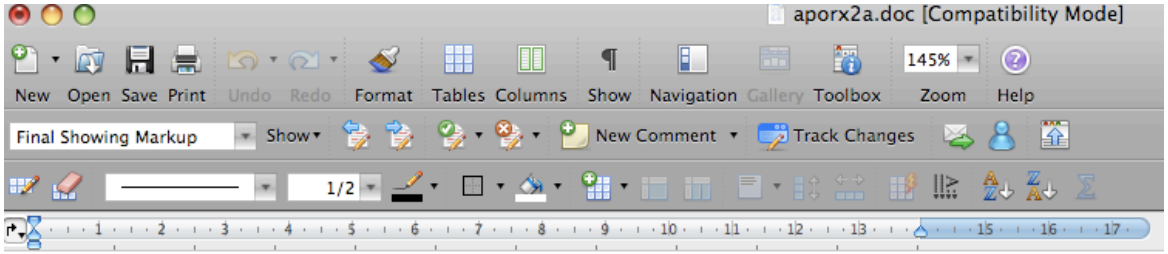
Gépi normalizálás

Morfológia

Felépítés

Lekérdezés

További feladatok



2. kéz

A, Á

1	2	3	4	5	6	7	8	9	10	11
079/07	a¹ nm	a	1000	000000	06	11	000	00	00	00
080/08	a	a	1000	000000	06	11	000	00	00	00
144/21	a	a	1000	000000	06	11	000	00	00	00
161/01	a	a	1000	000000	06	11	000	00	00	00
043/06	a² ne	a	2000	000000	13	11	000	00	00	00
	L. Függelék.									
115/11	Abiron	abironnak	0000	000000	03	11	000	00	03	01
113/21	Ábrahám	abrahamna	0003	000000	03	11	000	00	00	01

oldal	sor	betűhű	norm.	tő	elemzés
001	01	Mÿ	mi	mi	Pro.Nom_Gen
001	01	vronknac	urunknak	úr	N.PxP1.Dat_Gen
001	01	iesus	Jézus	Jézus	N:P.Nom
001	01	cristusnac	Krisztusnak	Krisztus	N:P.Dat_Gen
001	01	gyczeretÿre	dicséretire	dicséret	N.PxS3=i.Sub
001	02	es	és	és	C
001	02	gyczewsegere	dicsőségére	dicsőség	N.PxS3.Sub
001	02	es	és	és	C
001	02	my	mi	mi	Pro.Nom_Gen
001	02	atyancnak	atyánknak	atya	N.PxP1.Dat_Gen
001	03	bodog	boldog	boldog	Adj
001	03	ferencznek	Ferencnek	Ferenc	N:P.Dat_Gen

Anyaggyűjtés: nem szöveges formátumok

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

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

Szkennelés

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

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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* → *ć* ∼ L ∼ L’
3. keverék, pl. *cs* → *ch* ∼ *chy* ∼ *cyh* ∼ *c* ∼ *chi* ∼ *ch’* ∼ *cz* ∼ *ts* ∼ *ć* ∼ L ∼ L’ ∼ L*h* ∼ L*z*

Az OCR kiértékelése

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

nem törekszünk paleográfiai pontosságra:
az átirat szerkesztőjének konvencióit követjük

Példa

P. Balázs (1981)	*Jakab (2002)*
ıeſus	*iesus*
ſcent	*scent*

ahol egyedi indokkal mégis eltérünk ettől, azt külön jelezzük

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. $e + ‘ + \bar{} = \bar{e}$

UTF-8 kódolású sztenderd Unicode karakterek

az egész korpuszra kiterjedő szigorúan egységes formátum, DE: előfordulnak olyan régi karakterek, melyek a Unicode-ban nincsenek reprezentálva $\rightarrow$ helyettesítő karakter, pl. $\mathrm{L} \rightarrow \check{c}$

Prószéky-kódok

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a szövegeket Prószéky-kódokkal is tároljuk
betű-szám kombinációk
a Magyar Történeti Korpusz kódtábláját bővítettük ki

ékezet	kód	példa	kombináció
´	1	á	a1
¨	2	ö	o2
˝	3	ű	u3
‘	4	è	e4

Heterogén helyesírási rendszer

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,β* hangok bármelyikét

A normalizálás alapelvei 1.

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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
ymaduam	imádvám	imádva E/1.

A normalizálás alapelvei 2.

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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

Szavakra és mondatokra bontás

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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

A mondatokra bontás elvei

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a korabeli központozásra nem támaszkodhatunk → kézi mondatra bontás

az alárendelő tagmondatot nem választjuk el a főmondattól, a mellérendelőt igen

kétséges esetben inkább nem teszünk mondathatárt

A nevek egységesítése

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a bibliai tulajdonneveket a Szent István Társulati bibliafordítás alapján egységesen normalizáljuk

Példa

betűhű	*normalizált*
Rutthot	*Rutot*
Ruth	*Rut*
Rutnac	*Rutnak*

a SzIT-ben sem egységes névváltozatok közül a gyakoribbat használjuk

Gépi normalizálás

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a legjobb n normalizált alakot tartalmazó lista

a manuális annotáció redukálható a felkínált alakok közötti választásra

Példa

ygen (igen) $\rightarrow$

-10,8729908279143	*igén*
-11,3178857141749	*igen*
-11,5989613202567	*igény*
-13,4229320257043	*igyen**

Morfológiai elemzés és egyértelműsítés

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

félautomatikus normalizálás → *automatikus morfológiai elemzés* → *félautomatikus egyértelműsítés*

Humor – bővített lexikonnal és szabályhalmazzal

Szövegfeldolgozottsági szintek

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

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

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

Az annotációs szintek

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

a korpuszban minden egyes szövegszó mellett szerepelnek a következő adatok:

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

A korpusz anyaga

vertikális fájlformátum

kéz	könyv	oldal	fejezet	vers	betűhű	norm	ért	megj
1	Rut	4	2	8	Es	és		
1	Rut	4	2	8	monda	mondá		
1	Rut	4	2	8	Booz	Boász		
1	Rut	4	2	8	[Noėminèc]			FAIL
1	Rut	4	2	8	Rutnac	Rutnak:		
1	Rut	4	2	8				
1	Rut	4	2	8	Halgaſſad	hallgassad,		
1	Rut	4	2	8	leaṅom ·	leányom:		

Metaadatok

- **lókuszjelölők:** az adott szó helye az eredeti kódexben
  - kéz
  - bibliai fejezet- és versszámozás
- **értelmezés:** a normalizált alak mai magyarra való „fordítása”
- **megjegyzés:** TITLE, LANG{nyelv}, ADD, RECO, STRIKE, FAIL, FRAG

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: (3)*

A korpuszlekérdező felülete

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

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

A lekérdezés eredménye

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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, (szíve)**	mint	pávának.

[2] OMS - 9 - 1/130357

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

[3] OMS - 10 - 1/130354

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

Hogy keressünk rá arra, ami nincs ott?

a mai magyartól eltér az ómagyar névelőhasználat: sok helyen nincs névelő, ahol ma van

Példa

JokK 140.o.: Es azert ewkewztewk zent ferencz czudalatost gÿczerÿuala teremtewtt
[És azért őköztük Szent Ferenc csodálatost dicséri vala Teremtőt.]

megoldás: két olyan szóra keresünk rá, amelyek között várnánk a névelőt

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

További feladatok

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
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

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

Korpuszépítés ómagyar kódexekből
Simon Eszter
Bemutatás
Anyaggyűjtés
Feldolgozás
Szkennelés
OCR
Betűhű szöveg
Normalizálás
Gépi normalizálás
Morfológia
Felépítés
Lekérdezés
További feladatok

Köszönöm a figyelmet!