Magyar nyelvű történeti korpuszok

Simon Eszter

Debrecen, 2019. február 7.

MTA Nyelvtudományi Intézet

Az előadás vázlata

1. A történeti korpuszok jellemzői
2. A történeti szövegek feldolgozása
3. A korpuszok és amire használni lehet őket

A történeti korpuszok jellemzői

A történeti korpusz…

- szövegek vagy szövegrészletek véges elektronikus gyűjteménye, amely jól körülhatárolt és nyelvészetileg releváns kritériumok alapján lett válogatva, valamint legalább törekszik a reprezentativitásra;
- azzal a céllal készül, hogy reprezentálja egy nyelv régi állapotait és/vagy hogy tanulmányozni lehessen rajta a nyelv változásait;
- legalább egy generációnyit visszanyúlik a mai nyelvállapot előttre.

A történeti korpusz típusai

szinkrón	diakrón
Century of Prose Corpus (1680–1780)	Helsinki Corpus of English Texts (750–1710)
általános	specifikus
Helsinki Corpus of English Texts	Electronic Beowulf

Időkeret

Az időkeret

a történeti korpuszok esetében kiemelt fontosságú.
Kevesebb anyag áll rendelkezésre a régebbi korokból → még a szinkrón korpuszok is legalább egy évszázadot fognak át.

A méret

összefügg az időkerettel.
Minél nagyobb az időkeret, annál nagyobb a korpusz.

Korlátozó tényezők

A történeti korpuszok jellemzően kisebbek, mint a modernek.

- a történeti szövegek feldolgozása nagyobb kihívást jelent
- az anyag elérhetősége (történeti okok)
- az anyag elérhetősége (gyakorlati szempontok)

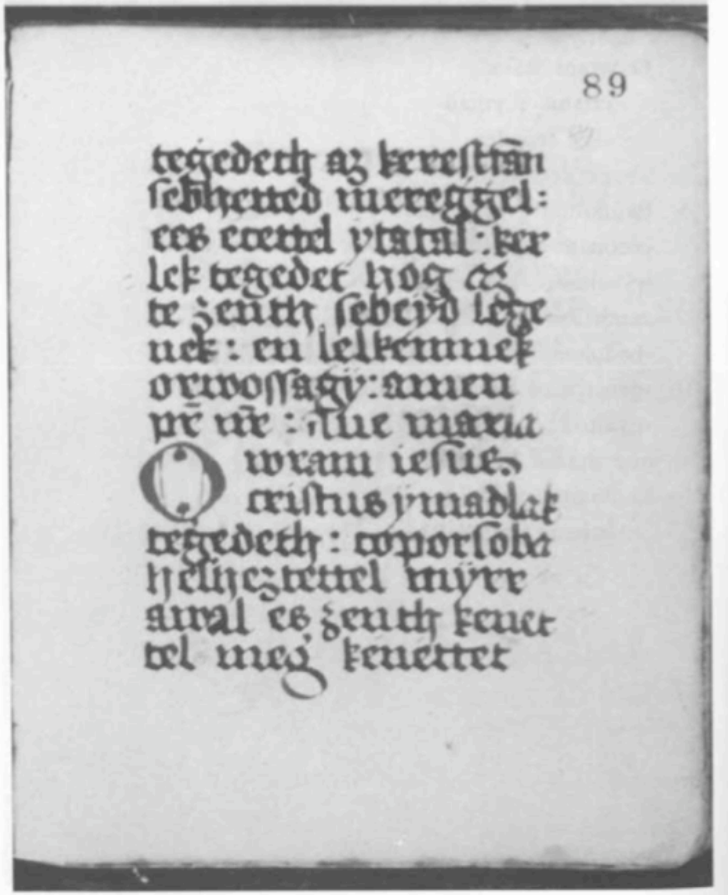
Reprezentativitás

nyelven kívüli tényezők

- a fennmaradt szövegek köre → csak egy random mintát szolgáltatnak a múltbeli nyelvről, nem reprezentálják a nyelv teljes változatosságát
- a beszélt nyelvi adatok hiánya
- a különféle műfajok, regiszterek aránya az írott nyelvváltozaton belül is kiegyensúlyozatlan → sok a vallási irodalom, kevés az informális, a sajtó, a tudományos stílus
- szociolingvisztikai kiegyensúlyozatlanság → a fennmaradt szövegek a társadalmi elit és az értelmiség nyelvét reprezentálják

A történeti szövegek feldolgozása

Kézzel írott kódexek, nyomtatott kiadások



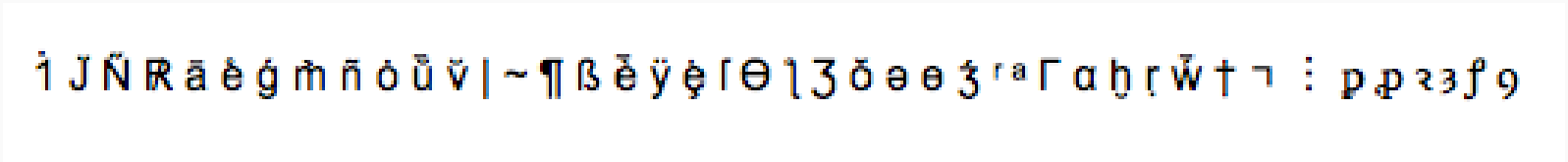
177
89r

tegedeth az kerestfan
sebhetted meregĝel :
ees ecettel ytatal : ker-
-lek tegedet hog az
te zenth sebeyd legé-
-nek : en lelkemnek
orwossagy : amen
pr nr : Aue maria
O wram iesus
cristus ymadlak
tegedeth : coporsoba
helheztettel myrr-
-awal es zenth kenet-
-tel meg kenettet

Speciális karakterek

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

Összesen: 156 karakter + ezek kombinációi



Heterogén helyesírás, normalizálás

a már nem létező jelenségeket megőrizni

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

a helyesírási esetlegességeket eltörölni

meden ∼ menden ∼ mendun ∼ mendē ∼ mendė ∼ miden ∼ minden ∼ mynden ∼ mű̈nden ∼ mű̈ndē ∼ mýden ∼ mýnden ∼ mýndew ∼ mÿnden ∼ mÿndon ∼ mēden ∼ mēdèn ∼ mēdē ∼ mēdėn ∼ mėnden ∼ ṁēden → minden

Nyelvi annotáció

az elérhető elemzők a modern nyelvállapotra készültek

→ a régi nyelvállapot elemzésére alkalmas eszközöket kell fejleszteni

→ az adaptáció nem triviális

→ több kézi ellenőrzést igényel

A korpuszok és amire használni lehet őket

Magyar nyelvű történeti korpuszok

1. Ómagyar Korpusz
   - MTA Nyelvtudományi Intézet
   - `http://omagyarkorpusz.nytud.hu`
   - 896–1526 & 1526–1772
2. Történeti Magánéleti Korpusz
   - MTA Nyelvtudományi Intézet
   - `http://tmk.nytud.hu/`
   - 1772 előtt
3. Magyar Történeti Szövegtár
   - MTA Nyelvtudományi Intézet
   - `http://www.nytud.hu/hhc/`
   - 1772 után

Egyéb források 1.

Számítógépes Nyelvtörténeti Adattár

a {Jókai-kódex, Birk-kódex, Guary-kódex, Apor-kódex, Festetics-kódex} ábécérendes adattára

`http://mnytud.arts.unideb.hu/sznyta.php`

Margit-legenda

a készülő kritikai szövegkiadás online változata

`http://deba.unideb.hu/deba/Margit-legenda_Szent_Margit_elete_1510/index.html`

Egyéb források 2.

Magyar Antikvakorpusz

válogatás a magyar nyelvű könyvnyomtatás első fél évszázadából (1527–1576)

`http://korpusz.ektf.hu/hu`

Sermones

a késő középkori és kora újkori magyarországi prédikációirodalom forrásszövegei

`http://sermones.elte.hu/`

A korpuszkeresők…

- segítségével elő tudunk állítani konkordancialistát és gyakoriségi listát is;
- alkalmasak arra, hogy történeti lexikológiai és szociolingvisztikai kutatásokhoz segítséget nyújtsanak;
- a morfológiai elemzést is tartalmazó korpuszok lehetőséget adnak történeti morfológiai vizsgálatok folytatására is, illetve bizonyos szintaktikai jelenségek is kutathatóak rajtuk.

Hipotézis

É. Kiss (2014), Gugán (2015, 2017)

kétféle tagadó szerkezet létezik a magyarban

- az egyenes szórendű (igekötő–tagadószó–ige, pl. *meg ne fogd*) → régebbi
- a fordított szórendű (tagadószó–ige-igekötő, pl. *ne fogd meg*) → újabb

hipotézis:

a fordított szórend is létezett a magyar nyelv korábbi szakaszaiban is, de csak a 19. századtól vált uralkodóvá

É. Kiss, K. (2014): A tagadó és a kérdő mondatok változásai. In: É. Kiss (szerk.): Magyar generatív történeti mondattan, 34–49. o. Akadémiai Kiadó, Budapest.
Gugán, K. (2015): És mégis: mozog? Tagadás és igemódosítók az ómagyarban és a középmagyarban. Általános Nyelvészeti Tanulmányok, 27:153–178.
Gugán, K. (2017): A magyar tagadó mondatok szórendje és a konstansráta-hipotézis. In: Nyelvelmélet és diakrónia 3., 91–110. o. Pázmány Péter Katolikus Egyetem BTK; Szt. István Társulat, Budapest; Piliscsaba

Vizsgálat

Kalivoda (2017)

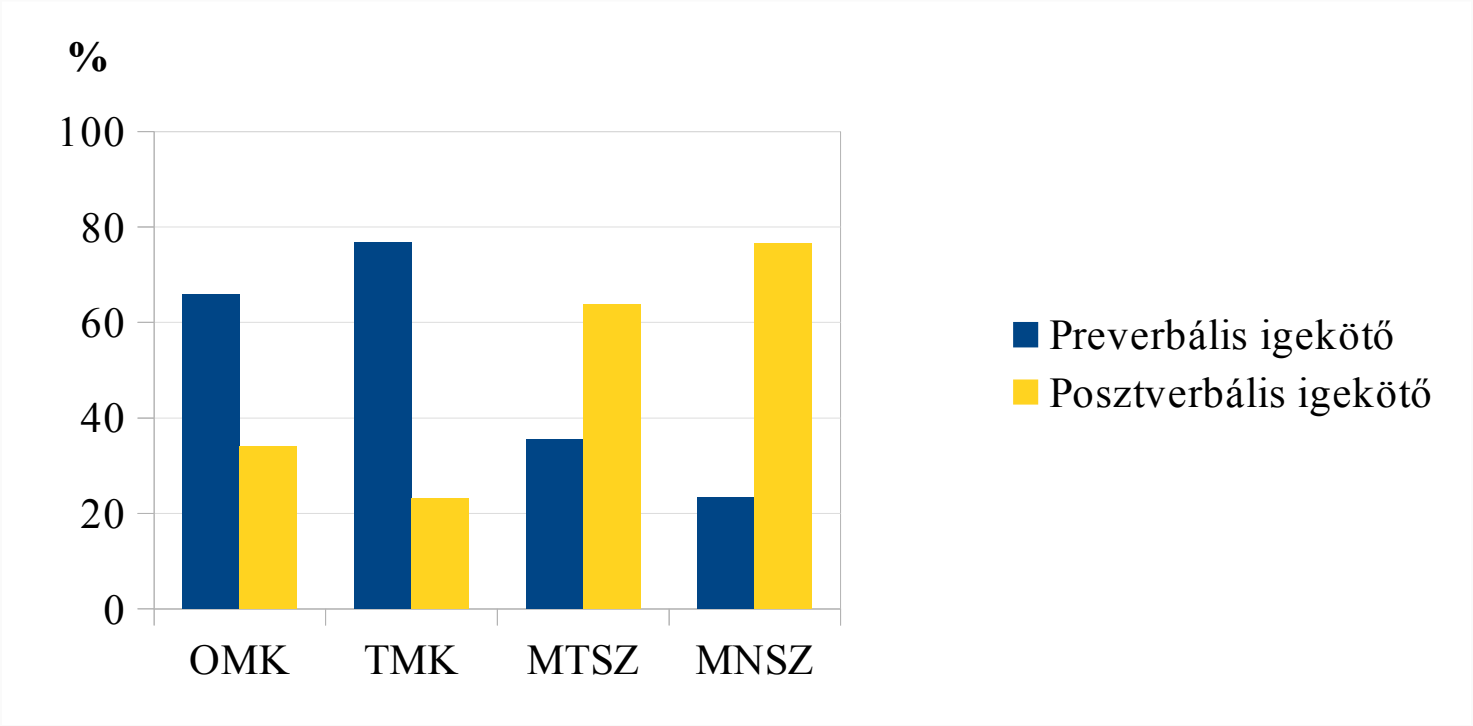
prototipikus igekötők (*meg*, *el*, *fel*, *ki*, *be*, *le*) szintaktikai viselkedését vizsgálta az ómagyar kortól napjainkig

a tagadó és a tiltó mondatokra jellemző preverbális (ige előtti) és a posztverbális (ige utáni) igekötők arányát vizsgálta

- Ómagyar Korpusz (ÓMK)
- Történeti Magánéleti Korpusz (TMK)
- Magyar Történeti Szövegtár (MTSz)
- Magyar Nemzeti Szövegtár (MNSz)

Kalivoda, Á. (2017): Prototipikus igekötők mondatbeli helye az ómagyar kortól napjainkig. Előadás a PPKE BTK Nyelvtudományi Doktori Iskolájának házi doktoranduszkonferenciáján.

Eredmények


%
100
80
60
40
20
0
OMK
TMK
MTSZ
MNSZ
Preverbális igekötő
Posztverbális igekötő

Javasolt olvasmányok

- Claudia Claridge: Historical corpora. In: Lüdeling, A. and Kytö, M. (eds.): Corpus Linguistics. An International Handbook. Walter de Gruyter, Berlin, 2008. 242–259.
- Matti Rissanen: Corpus linguistics and historical linguistics. In: Lüdeling, A. and Kytö, M. (eds.): Corpus Linguistics. An International Handbook. Walter de Gruyter, Berlin, 2008. 53–68.
- Anne Curzan, Ann Arbor: Historical corpus linguistics and evidence of language change. In: Lüdeling, A. and Kytö, M. (eds.): Corpus Linguistics. An International Handbook. Walter de Gruyter, Berlin, 2008. 1091–1109.
- Claire Bowern, Bethwyn Evans (eds.): The Routledge Handbook of Historical Linguistics. Routledge, London–New York, 2015.

Köszönöm a figyelmet!

simon.eszter@nytud.mta.hu

http://www.nytud.hu/oszt/korpusz/Simon_Eszter.html