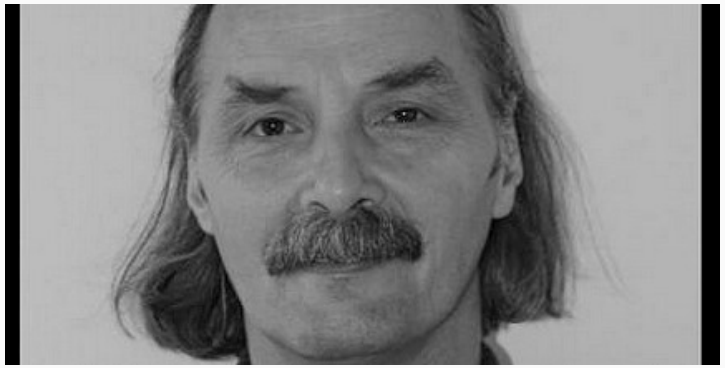
Magyar nyelvű történeti korpuszok

Simon Eszter

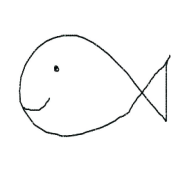
Pécs, 2019. május 2.

MTA Nyelvtudományi Intézet

Székely Gábor (1954–2019)



kala ∼ hal



Jókai-kódex (148,11-12)

„Veni ad me, frater lupe”

„ÿewÿ ydee, Attyamfya ffarkas”

Az előadás vázlata

1. Általános bevezetés
2. A történeti korpuszok jellemzői
3. A történeti szövegek feldolgozása
4. Korpuszok és egyéb források

Általános bevezetés

Szinkrónia és diakrónia

a szinkrón vagy leíró vizsgálat

- **tárgya:** a nyelv egyidejű állapota
- **célja annak megállapítása, hogy:** hogyan működik/ött a nyelv az adott időszakban

a diakrón vagy történeti vizsgálat

- **tárgya:** a nyelv időbeli egymásutánja
- **célja annak megállapítása, hogy:** mi változott, és mi nem változott

A magyar nyelv történetének korszakai

„előmagyar”		– i.e. 1000
ősmagyar		i.e. 1000 – i.sz. 896
ómagyar	korai ómagyar	896 – 1370
	kései ómagyar	1370 – 1526
középmagyar		1526 – 1772
újmagyar		1772 – 1920
újabb magyar		1920 –

A nyelvtörténet forrásai

1. **a nyelvemlékek:** *„a nyelvtörténeti vizsgálatok fundamentuma”*
2. **a mai nyelv:** *„a nyelvtörténet időrendben s átmenetileg utolsó láncszeme”*
3. **a rokon nyelvek:** *„alapvető fontosságúak a nyelvemléktelen kor nyelvi történéseinek a kikövetkeztetésében”*
4. **a kontaktusnyelvek:** *„lehetővé teszi a szókészleti és hangtani interferenciákon kívül a nyelv egyéb szintjein jelentkező egymásra hatások kimutatását”*

A történeti korpuszok jellemzői

Mi a korpusz? 1.

Kugler és Tolcsvai Nagy (2000)

„meghatározott szempontok alapján kiválasztott szövegmennyiség, amelyen a nyelvész vizsgálatát végzi”

Kugler Nóra és Tolcsvai Nagy Gábor (2000): Nyelvi fogalmak kisszótára. Korona, Budapest.

A főbb kérdések

- mennyiség
- nyelvészeti vizsgálatokra alkalmas
- reprezentativitás, a kiválasztás szempontjai
- tárolás módja: elektronikus
- tartalom: szegmentálás, annotáció, metaadatok

Mi a korpusz? 2.

Sinclair (2005)

„a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research”

J. Sinclair (2005): Corpus and Text – Basic Principles. In: M. Wynne (ed.): Developing Linguistic Corpora: a Guide to Good Practice. Oxbow Books, Oxford. 1–16.

Korpusz vs adatbázis

ID	betűhű	normalizált	tő	elemzés
1	Athyanak	Atyának	Atya	N:P.Dat_gen
2	ees	és	és	C
3	fýwnak	Fiúnak	Fiú	N:P.Dat_gen
4	ees	és	és	C
5	zenth leleknek	Szentléleknek	Szentlélek	N:P.Dat_gen
6	newe-@@ben	nevében	név	N.PxS3.Ine

ID	szerző	cím	év	hely	kiadó
1	Nietzsche	Az Antikrisztus	1993	Budapest	Ictus
2	Nietzsche	Ecce homo	1993	Budapest	Göncöl
3	Nietzsche	Bálványok alkonya	2004	Budapest	Holnap

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)

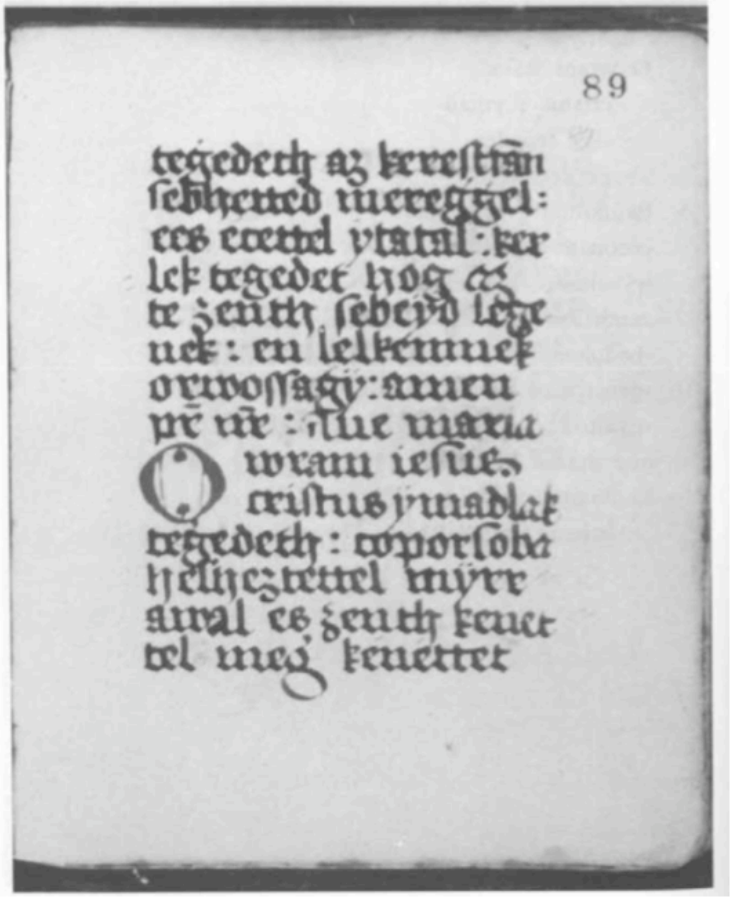
Reprezentativitás

nyelven kívüli tényezők

- a fennmaradt szövegek köre → csak egy véletlen 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

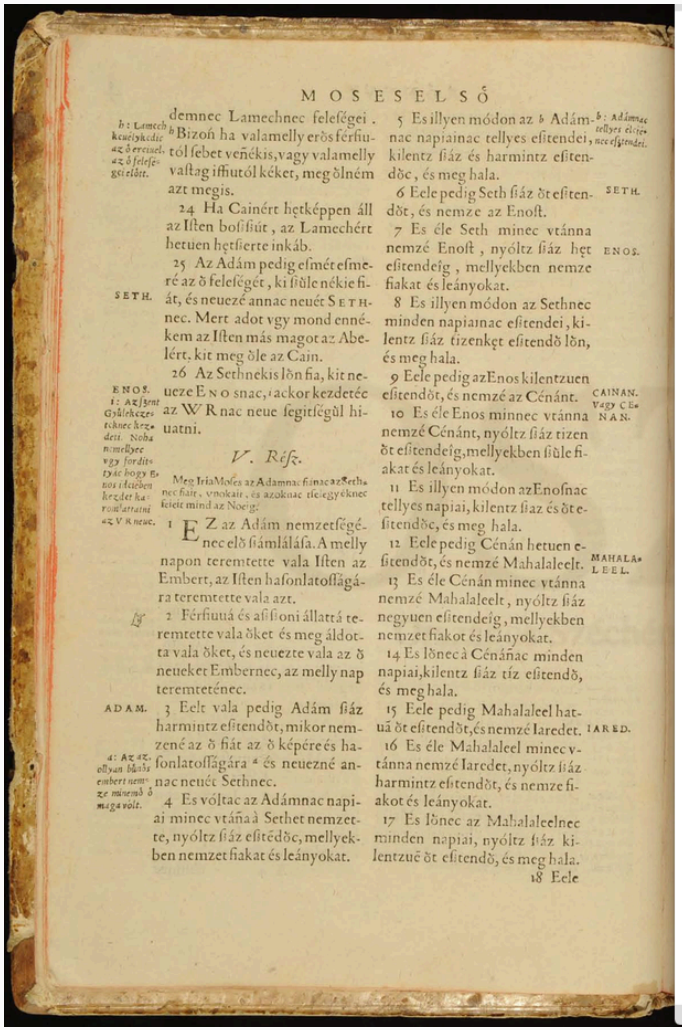
Kézzel írott kódexek, nyomtatott kiadások



177
89r

tegedeth az kerestfan
sebhetted mereggel :
ees écettel 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 mÿrr-
-awal es zenth kenet-
-tel meg kénéttet

Korai nyomtatványok, revideált változatok



MOSES ELSŐ

demnec Lamechnec feleſégei.
b Bizoń ha valamelly erös férfiu-
tól febet veńékis, vagy valamelly
vaſtag iſſiutól kéket, meg ölnem
azt megis.

24 Ha Cainért hętköppen áll
az Iſten boſiſiút, az Lamechért
hetuen hętſierte inkáb.

25 Az Adám pedig eſmét eſmé-
ré az ő feleſégét, ki ſiüle nékie fi-
át, és neuezé annac neuét SETH-
nec. Mert adot vgy mond enné-
kem az Iſten más magot az Abe-
lért, kit meg öle az Cain.

26 Az Sethnekis lőn fia, kit ne-
uezte ENOS nac, a kor kezdetéc
az WR nac neue ſegitſégül hi-
uatni.

b: Lamech keueiykedik az ő erecuel az ő feleſégei ellött.

SETH.

ENOS.

1: Az Jſten Gyülekezetknec kezdeti. Noha nemellyek vgy fordityác hogy Enos idejében kezdet kaualtatnai az VR neue.

V. Réſz.

Meg IriaMoſes az Adámnac ſiának az Sethnec fiait, vnokait, és azoknac ſiei egyéknec feleit mind az Noéig.

1 EZ az Adám nemzetſégénec elő ſiámláláſa. A melly napon teremtette vala Iſten az Embert, az Iſten haſonlatoſſágára teremtette vala azt.

2 Férfiuvá és aſiſioni állattá teremtette vala őket és meg áldotta vala őket, és neuezte vala az ő neueket Embernec, az melly nap teremteténec.

3 Eelt vala pedig Adám ſiáz harmintz eſtendőt, mikor nemzené az ő fiát az ő képére és haſonlatoſſágára [4] és neuezné annac neuét Sethnec.

4 Es vóltac az Adámnac napiai minec vtánna à Sethet nemzette, nyóltz ſiáz eſitédőc, mellyekben nemzet fiakat és leányokat.

ADAM.

4: Az az. ollyan féleb embert nemze minemő ő maga volt.

5 Es illyen módon az b Adámnac napiainac tellyes eſitendei, kilentz ſiáz és harmintz eſitendőc, és meg hala.

b: Adámnac tellyes eſitendei.

6 Eele pedig Seth ſiáz öt eſitendőt, és nemze az Enoſt.

SETH.

7 Es éle Seth minec vtánna nemzé Enoſt, nyóltz ſiáz hęt eſitendeſig, mellyekben nemze fiakat és leányokat.

ENOS.

8 Es illyen módon az Sethnec minden napiainac eſitendei, kilentz ſiáz tizenkęt eſitendő lőn, és meg hala.

9 Eele pedig azEnos kilentzuen eſitendőt, és nemzé az Cénánt.

CAINAN. Vagy CENAN.

10 Es éle Enos minnec vtánna nemzé Cénánt, nyóltz ſiáz tizen öt eſitendeſig, mellyekben ſiüle fiakat és leányokat.

11 Es illyen módon azEnoſnac tellyes napiai, kilentz ſiaz és öt eſitendőc, és meg hala.

12 Eele pedig Cénán hetuen eſitendőt, és nemzé Mahalaleelt.

MAHALALEEL.

13 Es éle Cénán minec vtánna nemzé Mahalaleelt, nyóltz ſiáz negyuen eſitendeſig, mellyekben nemzet fiakot és leányokat.

14 Es lőnec à Cénánac minden napiai, kilentz ſiáz tíz eſitendő, és meg hala.

15 Eele pedig Mahalaleel hatuā öt eſitendőt, és nemzé Iaredet.

IARED.

16 Es éle Mahalaleel minec vtánna nemzé Iaredet, nyóltz ſiáz harmintz eſitendőt, és nemze fiakot és leányokat.

17 Es lőnec az Mahalaleelnec minden napiai, nyóltz ſiáz kilentzuē öt eſitendő, és meg hala.

18 Eele

MÓZES I. KÖNYVE 4. 5. 9

feleségeinek: Oh Háda és Czilla, hallgassatok szómra, Lámekh feleségei, halljátok beszédem: embert öltem, mert megsebezett; ifjat öltem, mert megütött.

24. Ha * hétszeres a bosszú Kainért, hetvenhétszeres az Lámekhért. * vers 15.

Séth és Énós

25. Ádám pedig ismét ismeré az ő feleségét, és az szüle néki fiat, és nevezé annak nevét * Séthnek: mert adott *ugymond*, énnékem az Isten más magot Ábel helyett, kit megöle Kain. * rész 5,3.

26. Séthnek is született fia, és nevezé annak nevét Énósnak. Akkor kezdték segítségül hívni az Úrnak nevét.

Ádám nemzetsége a Séth ágán; a Noéig való pátriárkák
(V. ö. 1 Krón. 1,1—4.)

5 Ez az Ádám nemzetségének könyve. A mely napon teremté Isten az * embert, Isten hasonlatosságára teremté azt. * rész 1,26.9,6.

2. Férfiává * és asszonynyá teremté őket, és megáldá őket és nevezé az ő nevöket Ádámnak, a mely napon teremtetének. * rész 1,27.

3. Élt vala pedig Ádám száz harmincz esztendőt, és nemze *fiat* az ő képére és hasonlatosságára és nevezé annak nevét Séthnek.

4. És telének Ádám napjai, minekutánna * Séthet nemzette, nyolczszáz esztendőre, és nemze fiakat és leányokat. * 1 Krón. 1,1.

5. És lőn Ádám egész életének ideje kilenczszáz harmincz esztendő; és meghala.

6. Éle pedig Séth száz öt esztendőt, és nemzé Énóst.

7. És éle Séth, minekutánna Énóst nemzette, nyolczszáz hét esztendeig; és nemze fiakat és leányokat.

8. És lőn Séth egész életének ideje kilenczszáz tizenkét esztendő; és meghala.

9. Éle pedig Énós kilenczven esztendőt, és nemzé Kénánt.

10. És éle Énós, minekutánna Kénánt nemzette, nyolczszáz tizenöt esztendeig, és nemze fiakat és leányokat.

11. És lőn Énós egész életének ideje kilenczszáz öt esztendő; és meghala.

12. Éle pedig Kénán hetven esztendőt, és nemzé Mahalálélt.

13. És éle Kénán, minekutánna Mahalálélt nemzette, nyolczszáz negyven esztendeig; és nemze fiakat és leányokat.

14. És lőn Kénán egész életének ideje kilenczszáz tíz esztendő; és meghala.

15. Éle pedig Mahalálél hatvanöt esztendőt, és nemzé Járedet.

16. És éle Mahalálél, minekutánna Járedet nemzette, nyolczszáz harmincz esztendeig, és nemze fiakat és leányokat.

17. És lőn Mahalálél egész életének ideje nyolczszáz kilenczvenöt esztendő; és meghala.

18. Éle pedig Járed száz hatvankét esztendőt, és nemzé Énókhot.

19. És éle Járed, minekutánna Énókhot nemzette, nyolczszáz esztendőt; és nemze fiakat és leányokat.

20. És lőn Járed egész életének ideje kilenczszáz hatvankét esztendő; és meghala.

21. Éle pedig Énókh hatvanöt esztendőt, és nemzé Methuséláht.

22. És járt * Énókh az Istennel, minekutánna Methuséláht nemzette, háromszáz esztendeig; és nemze fiakat és leányokat. * rész 6,9.17,1.

23. És lőn Énókh egész életének ideje háromszáz hatvanöt esztendő.

24. **És mivel Énókh Istennel járt * vala; eltűnék, mert Isten magához vevé.** * Zsid. 11,5.

25. Éle pedig Methusélah száz nyolczvanhét esztendőt és nemzé Lámekhet.

26. És éle Methusélah, minekutánna Lámekhet nemzette, hétszáz nyolczvankét esztendőt; és nemze fiakat és leányokat.

27. És lőn Methusélah egész életének ideje kilenczszáz hatvankilencz esztendő; és meghala.

28. Éle pedig Lámekh száz nyolczvankét esztendőt, és nemze fiat.

29. És nevezé azt Noénak, mondván: Ez vigasztal meg minket munkálkodásunkban s kezünk terhes fáradozásában e földön, melyet megátkozott az Úr.

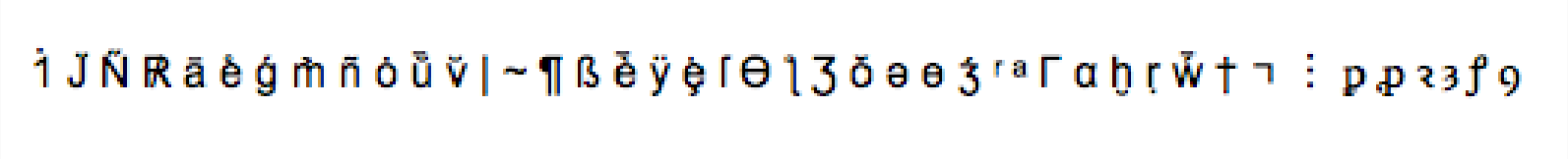
30. És éle Lámekh, minekutánna Noét nemzette, ötszáz kilenczvenöt esztendőt; és nemze fiakat és leányokat.

Rossz fizikai állapot

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 ∼ műnden ∼ műndē ∼ 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

Korpuszok és egyéb források

A nyelvtörténet forrásai

1. **a nyelvemlékek:** *„a nyelvtörténeti vizsgálatok fundamentuma”*
2. **a mai nyelv:** *„a nyelvtörténet időrendben s átmenetileg utolsó láncszeme”*
3. **a rokon nyelvek:** *„alapvető fontosságúak a nyelvemléktelen kor nyelvi történéseinek a kikövetkeztetésében”*
4. **a kontaktusnyelvek:** *„lehetővé teszi a szókészleti és hangtani interferenciákon kívül a nyelv egyéb szintjein jelentkező egymásra hatások kimutatását”*

A nyelvemlékek 1.

Magyar nyelvemlékek – OSZK

- OSZK
- `http://nyelvemlekek.oszk.hu`
- 1055–1526

KlimoTheca

- PTE Könyvtár
- `https://kt.lib.pte.hu/cgi-bin/kt.cgi`

OSZK Digitális Könyvtár

- OSZK
- `http://oszkdk.oszk.hu`

A nyelvemlékek 2.

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`

A nyelvemlékek 3.

Sermones

a késő középkori és kora újkori magyarországi prédikációirodalom forrásszövegei

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

Magyar Antikvakorpusz

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

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

A nyelvemlékek 4.

Ómagyar Korpusz

- MTA Nyelvtudományi Intézet
- `http://omagyarkorpusz.nytud.hu`
- 896–1526 & 1526–1772

Történeti Magánéleti Korpusz

- MTA Nyelvtudományi Intézet
- `http://tmk.nytud.hu/`
- 1772 előtt

Magyar Történeti Szövegtár

- MTA Nyelvtudományi Intézet
- `http://www.nytud.hu/hhc/`
- 1772 után

A mai nyelv

Magyar Nemzeti Szövegtár

- MTA Nyelvtudományi Intézet
- `http://clara.nytud.hu/mnsz2-dev/`
- 20. század vége – 21. század eleje

A rokon nyelvek

Uralonet – Uráli Etimológiai Adatbázis

- MTA Nyelvtudományi Intézet
- `http://uralonet.nytud.hu`

OUDB – Ob-Ugric Database

- Ludwig Maximilian University of Munich
- `http://www.oudb.gwi.uni-muenchen.de/`

A kontaktusnyelvek

TESz – Történeti-Etimológiai Szótár

- Arcanum
- `https://adtplus.arcanum.hu/hu/collection/MNYTESZ/`

UraLUID – Uralic Languages Under the Influence Database

- MTA Nyelvtudományi Intézet
- `http://www.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html`

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)

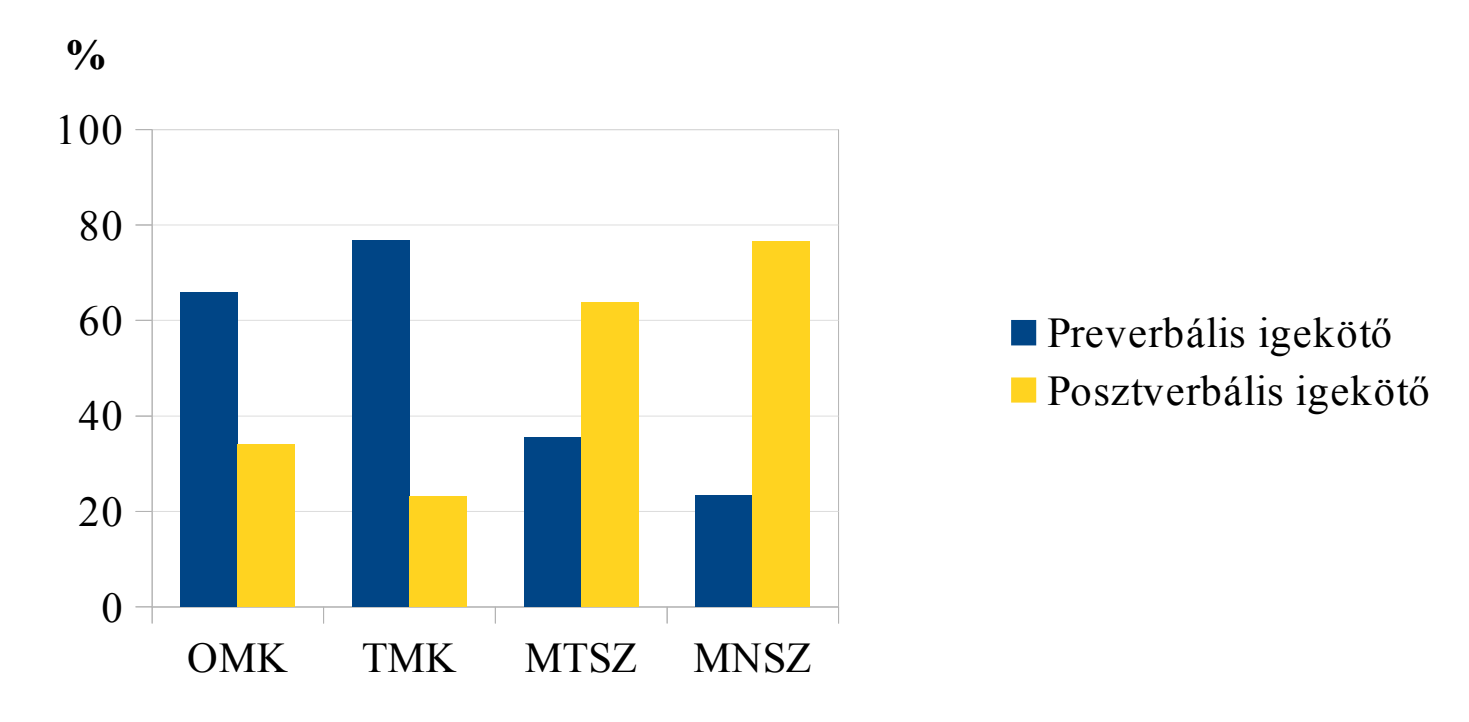
prototipikus igekötők (*meg, el, fel, ki, be, le*) szintaktikai viselkedését vizsgálta az ómagyar kortól napjainkig

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