Korpuszépítés

Nyelv és informatika – Pécs, 2022/23 tavasz
5. óra

Simon Eszter – Vadász Noémi

2023. április 1.

Tartalom

NYTK-NerKor

Motiváció

NER: Named Entity Recognition

- a felügyelt gépi tanuláson alapuló rendszerek nagy mennyiségű gold standard adatot igényelnek → így van ez a magyar NER-ben is
- a gold standard adat előállítása drága és erőforrásigényes → kevés van, azok pedig erősen domainspecifikusak, korlátozottak méretükben és hozzáférhetőségükben

Magyar NER korpuszok:

- Szeged NER korpusz: gold, kb. 200.000, hírek
- Criminal NE korpusz: gold, kb. 500.000, hírek
- hunNERwiki korpusz: silver, kb. 19M, wiki
- NYTK-NerKor korpusz

Az NYTK-NerKor korpusz

- 1 millió szavas
- gold standard
- kiegyensúlyozott válogatás 5 domainből (kb. 200.000 szó/domain):
  - szépirodalom
  - jogi szövegek
  - vegyes webes szövegek
  - hírek
  - Wikipédia-cikkek
- gold standard morfológiai elemzés a korpusz egyötödén
- CC-BY-SA 4.0 licenc alatt elérhető: https://github.com/nytud/NYTK-NerKor

Nemzetközi sztenderdek

- adatformátum: CoNLL-U Plus
- NE címkekészlet: CoNLL2002 (*PER*, *ORG*, *LOC*, *MISC*)
- NE címkeprefixálás: IOB2 (CoNLL2002)
- morfológiai elemzés: Universal Dependencies v2

A korpusz további tulajdonságai

- a korpusz előfeldolgozása az *emtsv*-vel történt
- további morfológiai elemzés az *emMorph* címkekészletében
- hivatalos train–devel–test vágás: 80%–10%–10% → kiegyensúlyozott: minden műfaj, forrás és morfológiai annotáltság ugyanilyen arányban van
- összehasonlításképpen: az OntoNotes 5.0 angol nyelvű része kb. 1,5 millió tokent tartalmaz

NE annotálás

név: egyedi referenciával bír, a világ egy egyedi entitására utal

- *Kosztolányi Dezső*
- *Szilas Menti Mezőgazdasági Termelőszövetkezet*
- *United Nations Educational, Scientific and Cultural Organization*
- *Déli-Shetland-szk.*
- *IBM*
- *Kiss János altábornagy utca*
- *Műegyetem*
- *The Coca-Cola Co.*
- *Kovács Pistike*

Az annotálás alapelvei I.

- Csak tulajdonneveket annotálunk. Nem annotálunk olyan frázisokat, amelyek ugyan a világnak valamely egyedi részére utalnak, de nem tulajdonnévvel. Például a *József Attila Gimnázium* annotálandó, de a szövegben szereplő az *a suli* frázis nem, hiába derül ki a szövegből, hogy melyik iskolára utal.
- A nevek nem kompozicionálisak. Mivel a nevek jelölete nem a név részeinek a jelöletéből áll össze, ezért a neveket nem bonthatjuk részekre az annotáláskor. Például a *Kossuth Lajos utca* egy földrajzi névként jelölendő, hiába van benne egy személynév. Ebből az következik, hogy mindig a leghosszabb nevet (a legkülsőbbet) jelöljük a jelölhetők közül.
- Nem annotálunk egymást átfedő vagy egymásba ágyazott neveket. Vagyis minden annotációnak be kell fejeződnie, mielőtt egy másik elkezdődik.

Az annotálás alapelvei II.

- A *tag-for-meaning* elvét követjük. Vagyis egy nevet mindig az aktuális kontextusnak megfelelő referenciája alapján annotálunk.
- Ha az azonosított név ragozott formában szerepel a szövegben, a raggal együtt, a teljes alakot annotáljuk.
- A nevek képzett alakjait nem jelöljük. Nem annotálandók tehát az olyanok, mint *magyarországi, fideszes, petőfieskedő*.
- Ha a név összetétel előtagja, és az összetétel alaptagja köznév, például *Horn-kormány, Tilos Rádió-hallgatók, TA-vezérigazgató*, akkor nem annotálandók névként.
- A névhez nem tartozik hozzá az esetleg előtte álló névelő. Kivétel az az eset, amikor a határozott névelő része a névnek, például *The Hague, The Times*.

Az annotálás alapelvei III.

- A név rövidítése (akronim, mozaikszó, monogram) is névként annotálandó.
- A szöveghez a névannotálás során nem nyúlunk hozzá, vagyis nem javítjuk ki a helyesírási hibákat, nem vonunk egybe különírt szavakat, és nem választunk szét egybeírtakat. Ha valamilyen éktelen hibát látunk az aktuálisan címkézendő névvel kapcsolatban, akkor azt külön fel kell jegyezni. Ez a szépirodalmi szövegekre nem vonatkozik: ott az az elv, hogy amit a szerző leírt, az sérthetetlen.

NE Kategóriák

- PERSON: Valós és kitalált személyek neve, becenevek, művésznevek, álnevek. Ide tartoznak a kisebb, kevésbé strukturált embercsoportok, közösségek is.
- ORGANIZATION: Olyan csoportok nevei, amelyek valamilyen szervezett struktúrával rendelkeznek, mint például intézmények, vállalatok, kormányzati hivatalok, sportcsapatok, múzeumok, egyetemek.
- LOCATION: Földrajzilag vagy politikailag definiált helyek nevei, úgymint városok, országok, hegyek, völgyek stb. Idetartoznak az emberalkotta építmények is, mint a repterek, utak, gyárak, épületek stb.
- MISC: A felsorolt típusok egyikébe sem tartozó nevek.

Azóta...

2021-es megjelenése óta:

- sokan használják
- a Stanza (Qi et al., 2020) magyar nyelvű NER modulja
- a HuSpaCy (Orosz et al., 2022) magyar nyelvű NER modulja
- kiértékelések: kellően nagy és heterogén $\rightarrow$ a rajta tanított rendszerek rendelkeznek azzal az általánosító képességgel, amivel eddig nem
- további munkálatok

KorKor

KorKor

- többrétegű, kézzel annotált korpusz
- morfológiai címke és tő kétféle címkekészlettel, dependenciaelemzés, zéró igék és zérónévmások, anaforikus- és koreferenciakapcsolatok
- kétféle fájlformátum
- a korpusz, a munkafolyamat és a felhasznált eszközök elérhetőek itt: https://github.com/vadno/korkor_pilot
- CC-BY-4.0. licenc

	dokumentum	token (*conllup*)	token (*xtsv*)
huwiki	62	16,739	18,262
globv	32	7,760	8,799
TOTAL	94	24,499	26,581

A munkafolyamat

1. szöveggyűjtés
2. *emtsv* elemzés (*emToken*, *emMorph*, and *emTag*)
3. formátumkonverzió
4. kézi ellenőrzés (Google Spreadsheets)
5. formátumkonverzió
6. *emtsv* elemzés (*emDep*)
7. formátumkonverzió (*emCoNLL*)
8. kézi ellenőrzés (WebAnno)
9. a zérólétigék és az elliptált igén kézi beillesztése (plain text editor)
10. zérónévmásbeszúrás (*emZero*)
11. névmási anafora beillesztése (saját szkript)
12. kézi ellenőrzés és koreferenciaannotálás (Google Spreadsheets)
13. formátumkonverzió

A zérónévmások beillesztése

- a magyar pro-drop nyelv, az ige alanya, tárgya és a birtok birtokosa lehet droppolva
- egy szabályalapú szkript illesztette be ezeket
- a szabályok a többi nyelvi annotáción (tő, morfológiai címke, dependenciaelemzés):
  - alany, ha a függőségi fában az igéhez nem tartozott alany;
  - tárgy, ha egy határozott ragozású igéhez nem tartozott tárgy a függőségi fában;
  - birtokos, ha egy birtokhoz nem tartozott birtokos a függőségi fában;
  - alany a ragozott vagy a ragozatlan infinitívuszhoz
- plusz ágak jelennek meg a függőségi fában
- a beillesztett névmások morfológiai jegyei (szám, személy) az ige vagy a tárgy morfológiai címkéje alapján kiszámolható
- 867 zéró alanyt, 101 zéró tárgyat és 379 zéró birtokost illesztettünk be

A névmási anaforikus kapcsolatok beillesztése

- egy szabályalapú szkript illesztette be
- a szkript megkeresi a személyes névmásokat
- a szabályok a szófaj, a morfológiai jegyek és a szintaktikai információk alapján működnek
- antecedenst keres a névmásoknak
- pl. ha az ige alanya zéró és az ige ragozása megegyezik az előző tagmondat igéével, akkor a zéró alany antecedense megegyezik az előző tagmondat alanyának antecedensével

Kézi javítás és koreferenciaannotáció

- 4 nyelvészhallgató ellenőrizte és javította a beillesztett zérónévmásokat és az anaforikus kapcsolatokat, valamint a annotálta a koreferenciakapcsolatokat
- Google Spreadsheets feltételes formázásokkal
- anafora típusok: személyes (**prs**), mutató (**dem**), kölcsönös (**recip**), visszaható (**refl**), vonatkozó (**rel**), birtokos (**poss**)

általános

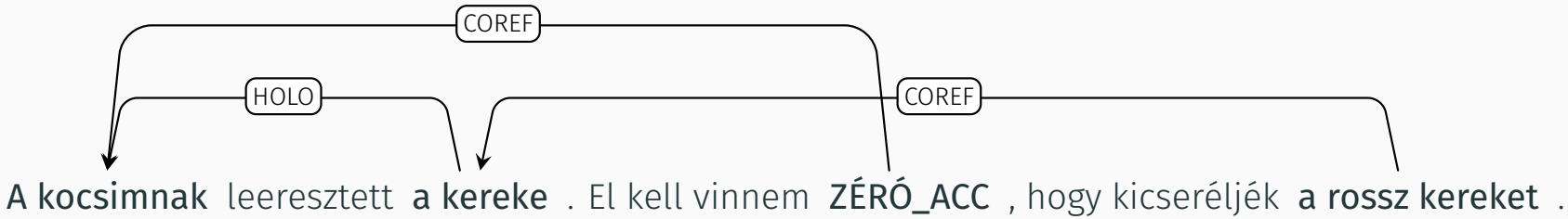
a Kínai Kommunista Párt egyik volt vezetője, akit hazaárulás miatt **elítéltek**

beszélő és címzett

A születésnapi ajándékoknak is nagyon **örülünk**, ha **szeretnéd** támogatni a munkánkat, **küldj nekünk** adományt, vagy **vegyél** egyet az NSA-s karácsonyi **üdvözlőlapjaink** közül, amelyet a Creative Time-nál dolgozó **barátaink** terveztek.

Koreferencia és a szerteágazó kapcsolatok

- koreferenciatípusok: két elemnek ugyanaz a referense (**coref**), rész-egész napcsolat áll fenn a két elem között (**holo**)
- szerteágazó kapcsolatok: a koreferenciakapcsolat csak testes elemek között lehet, a névmás antecedense lehet droppolt névmás is.



További nehézségek

- A referens állapota megváltozik. Vajon a holttest koreferens az emberrel?
  *Három hónap telt el az **újságíró házaspár**, Sagar Sarwar és felesége, Meherun Runi meggyilkolása óta. A **holttesteket** már exhumálták is, hogy megismételjék a boncolást.*
- split antecedens
  ***Papyrus** bátor és megmenti **Thèti-Chèri-t**. **A két egymásra lelt barát** küldetést kap az istenektől, hogy védelmezzék meg a fáraót.*

MoMa korpusz

MoMa korpusz

- moldvai magyar nyelvjárások
- Tánczos Vilmos kolozsvári néprajzkutató moldvai gyűjtései
- minden olyan moldvai települést felkeresett, ahol azt feltételezte, hogy tetten érhetők még addig nem dokumentált archaikus népi imádságok, így értékes anyag gyűlt fel a magyar e kevéssé dokumentált változatairól
- előzmény: https://nlp.nytud.hu/csango/
- eddig nem létezett ezeknek a nyelvjárásoknak a kutatására alkalmas, megfelelő pontossággal lejegyezett és nyelvi annotációval ellátott korpusza
- jelenleg 63 ezer token (30 interjú 21 kutatópontról)
- augusztusra várhatóan 25 kutatópontról 60 interjú lesz

Munkafolyamat

- lejegyzés a magyar egyezményes hangjelölés standardjával
- automatikusan egyszerűsített átirat
- a hangfelvételek össze vannak kapcsolva a szöveggel
- normalizálás
- morfológiai elemzés és tövesítés (az ómagyar elemző reszelgetésével)
- SketchEngine lekérdezőfelület
- nemcsak lekérdezőfelületen keresztüli, hanem teljes elérés

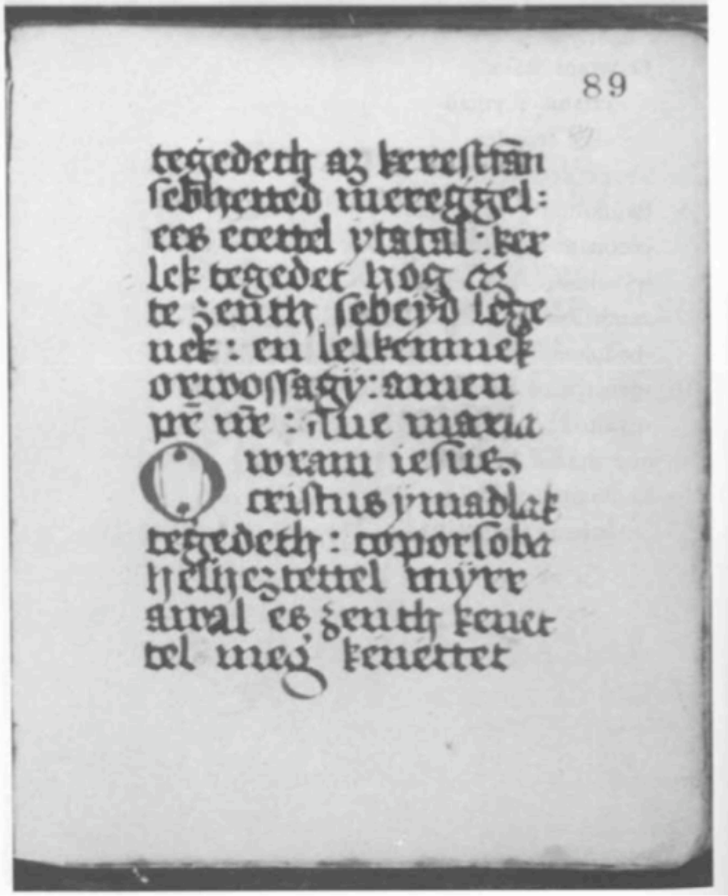
Kutatópontok

Ómagyar Korpusz

Ómagyar Korpusz (MGTSZ)

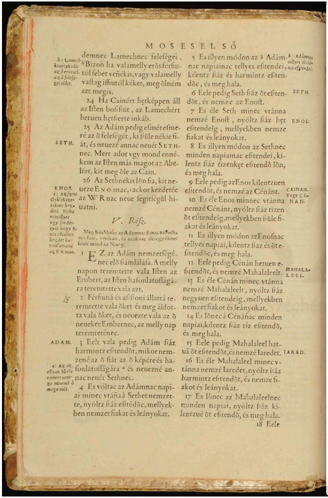
- 2009-2013, 2015-2018
- annotált korpusz, amely tartalmazza az összes fennmaradt ómagyar kori (896–1526) és néhány középmagyar kori (1526–1772) szövegemléket, és amely nyelvészetileg releváns kérdésekre tud választ szolgáltatni
- 3,2 millió szövegszó: 47 ómagyar kódex, 24 rövidebb ómagyar szövegemlék, 244 misszilis és 5 középmagyar kori bibliafordítás
- http://omagyarkorpusz.nytud.hu

Kézzel írott kódexek, nyomtatott kiadások



tegedeth az kerestfan
sebhetted meregĝel :
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
tegedeth : coporsoba
helheztettel mÿrr-
-awal es zenth kenet-
-tel meg kénéttet

Korai nyomtatványok, revideált változatok



MOSES ELSŐ

demnec Lamechnec feleſégei. [b] Bizoń ha valamelly erős férfiutól ſebet veńékis, vagy valamelly vaſtag iffiutól kéket, meg ölném azt megis.

24 Ha Cainért hętképpen áll az Iſten boſſiút, az Lamechért hetuen hętſierte inkáb.

25 Az Adám pedig eſmét eſmeré az ő feleſégét, ki ſiüle nékie fiát, és neuezé annac neuét SETHnec. Mert adot vgy mond ennékem az Iſten más magot az Abelért. kit meg öle az Cain.

26 Az Sethnekis lön fia, kit neueze ENOSnac, [i] ackor kezdetéc az WRnac neue ſegitſégül hiuatni.

b: Lamech keuélykedic az ő erciuel, az ő feleſégei elött.

SETH.

ENOS.

i: Az ſzent Gyülekezeteknec kezdeti. Noha némellyec vgy fordittyác hogy Enos idejében kezdet haromlattatni az VR neue.

V. Réſz.

Meg Iria Moſes az Adamnac fiánac az Sethnec fiait, vnokait, és azoknac tſelegyéknec feleit mind az Noeig.

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

2 Férfiuuá és aſſioni állattá teremtette vala őket és meg áldotta vala őket, és neuezte vala az ő neueket Embernec, az melly nap teremtetének.

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

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

ADAM.

a: Az az, ollyan bűnös embert nemze minemő ő maga vólt.

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

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

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

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

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

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

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

12 Eele pedig Cénán hetuen eſitendőt, és nemzé Mahalaleelt.

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

14 Es lönec à Cénáńac minden napiai, kilentz ſiáz tíz eſitendő, és meg hala.

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

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

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

18 Eele

b: Adámnac tellyes életének eſztendei.

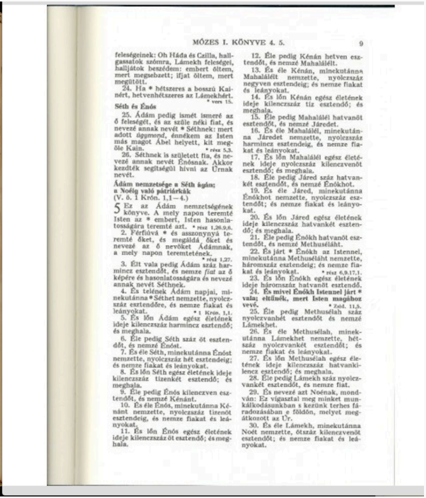
SETH.

ENOS.

CAINAN. Vagy CENAN.

MAHALALEEL.

IARED.



MÓZES I. KÖNYVE 4. 5. 9

feleségeinek: Oh Háda és Czilla, hallgassatok szómra, Lámekh feleségei, halljátok beszédem: embert öltem, mert megsebzett; 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 *úgymond*, 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. Férfiúvá * 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 vígasztal 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.

Érvek a szerkesztett kiadások...

...mellett	...ellen
• könnyen elérhetőek • a szerkesztői döntéseket már meghozták • könnyebben szkennelhető vagy begépelhető	• a szerkesztést nem nyelvészek végezték • a meghozott nyelvészeti döntések nincsenek kellően dokumentálva • nem alkalmasak további nyelvészeti vizsgálatokra • copyright

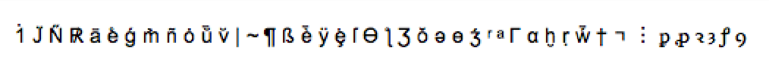
Kiút a dilemmából: szerkesztett kiadásokból kiindulni, de mindent ellenőrizni az eredeti verzióban.

Rossz fizikai állapot

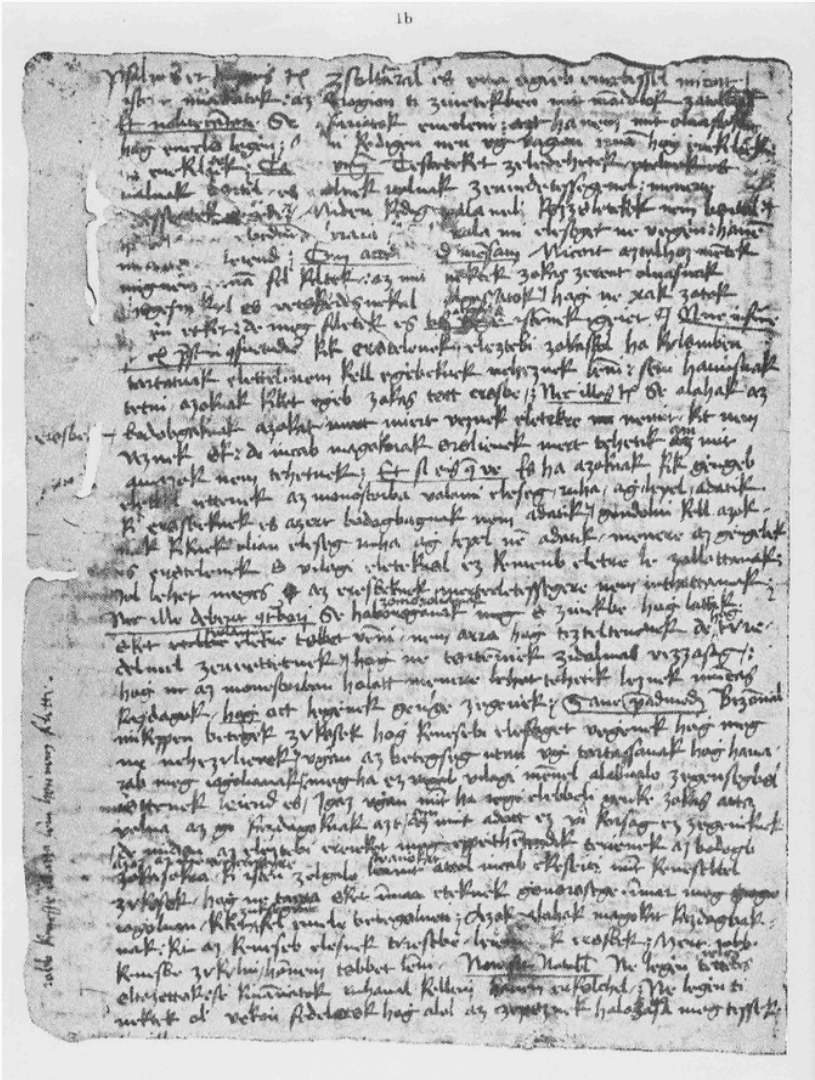
Speciális karakterek

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

Összesen: 156 karakter + ezek kombinációi



Betoldások, javítások, etc.



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

Megoldás: az eredeti és a normalizált verziót is tartalmazza a korpusz.

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

Uralic Languages Under the Influence Database

A projekt

- *Az uráli nyelvek mondattanának változása aszimmetrikus kontaktushelyzetben*
- 2016. február – 2017. július
- MTA Nyelvtudományi Intézet
- interdiszciplináris csapat: kutatók a finnugor, a nyelvtechnológiai és az elméleti nyelvészeti osztályról
- projektvezető: É. Kiss Katalin
- a vizsgált nyelvek: udmurt, tundrai nyenyec, szinjai és szurguti hanti
- célja egy pilot adatbázis építése volt:
  - 4000 token/kor/nyelv
  - IPA-átirat
  - teljes morfológiai elemzés
  - angol fordítás

A vizsgált nyelvek

Udmurt:

- írott (5) kategória: napi szinten használják, és létezik egy sztenderd irodalmi változata, de az nem annyira terjedt el
- Udmurtia egyik hivatalos nyelve
- vannak udmurt blogok, napi szinten keletkezik elektronikus udmurt szöveg
- udmurt Wikipédia (1000+ cikk)

Többiek:

- veszélyeztetett (6b) kategória: csak informális körben használják, alacsony presztízs
- nincs Wikipédia
- nincs napi sajtó, nem keletkezik elektronikus szöveg

A korpuszépítés kritériumai

- **elsődleges adatok:** olyan kommunikációs eseményekből származó nyelvi adatok, amelyek a hétköznapi nyelvhasználatot tükrözik
- **teljességre törekvés:**
  - minél több társadalmi osztályt, kort, nemet, műfajt és dialektust reprezentáljunk
  - metaadatok
  - az eredeti felvétel megőrzése, hogy a lejegyzések ellenőrizhetők legyenek
- **egységesség és összehasonlíthatóság:** nemzetközi sztenderdek
  - Unicode
  - IPA
  - Leipzig Glossing Rules
  - UTF-8 kódolású tsv fájlok

A korpusz felépítése

YRK Hajdú:	jā	mīdaxana	amkerta	jaŋkūwi
YRK Mus:	ja	midaxana	amkerta	jaŋkuwi
YRK IPA:	ja	mi:daxana	ămkerta	jăŋkuwi
YRK cirill:	я	мыдахана	амкэрта	яңкувы
lemma:	я	мы	ңамгэ	яңгось
szófaj:	N	Ptcp	Pron.neg	V
glossza:	earth	create.IPFV.PTCP.LOC	what.CONC	neg.EX.INFER

ENG:	when the earth was created, there was nothing
GER:	zur zeit der erschaffung der erde gab es nichts
HUN:	a föld teremtésének idején nem volt semmi

További néznivalók

http://archive.nytud.hu/oszt/elmnyelv/urali/adatbazisok.html

Simon (2017)

Simon and Mus (2017)

References

Orosz, Gy., Szántó, Zs., Berkecz, P., Szabó, G., and Farkas, R. (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. In *XVIII. Magyar Számítógépes Nyelvészeti Konferencia*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Simon, E. (2017). Négy hatás alatt álló nyelv – korpuszépítés kis uráli nyelvekre. In Vincze, V., editor, *XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017)*, page 263–274, Szeged.

Hivatkozások ii

Simon, E. and Mus, N. (2017). Languages under the influence: Building a database of uralic languages. In M. Tyers, F., Riessler, M., Pirinen, T. A., and Trosterud, T., editors, *Proceedings of the Third International Workshop on Computational Linguistics for Uralic Languages*, page 10–24, Saint Petersburg.