dHUpla: a kézirattól az automatikus nyelvfeldolgozásig

Szűcs Kata Ágnes – Simon Eszter
OSZK DBK
2023.08.31.

Digitális forráskiadás

Eddig publikálatlan források közzététele digitális szövegkiadás formájában.

https://dhupla.hu/page/adhuplarol

dhupla

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

STAGING SERVER

dhupla

digital humanities platform /// digitális bölcsészeti platform

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

Demo

Móricz Zsigmond – Holics Janka (1915-11-18)

Metaadatok | Megjegyzések

A feladó

neve:

Móricz Zsigmond kriegsbericht-erstatter*

címe:

Tp 39

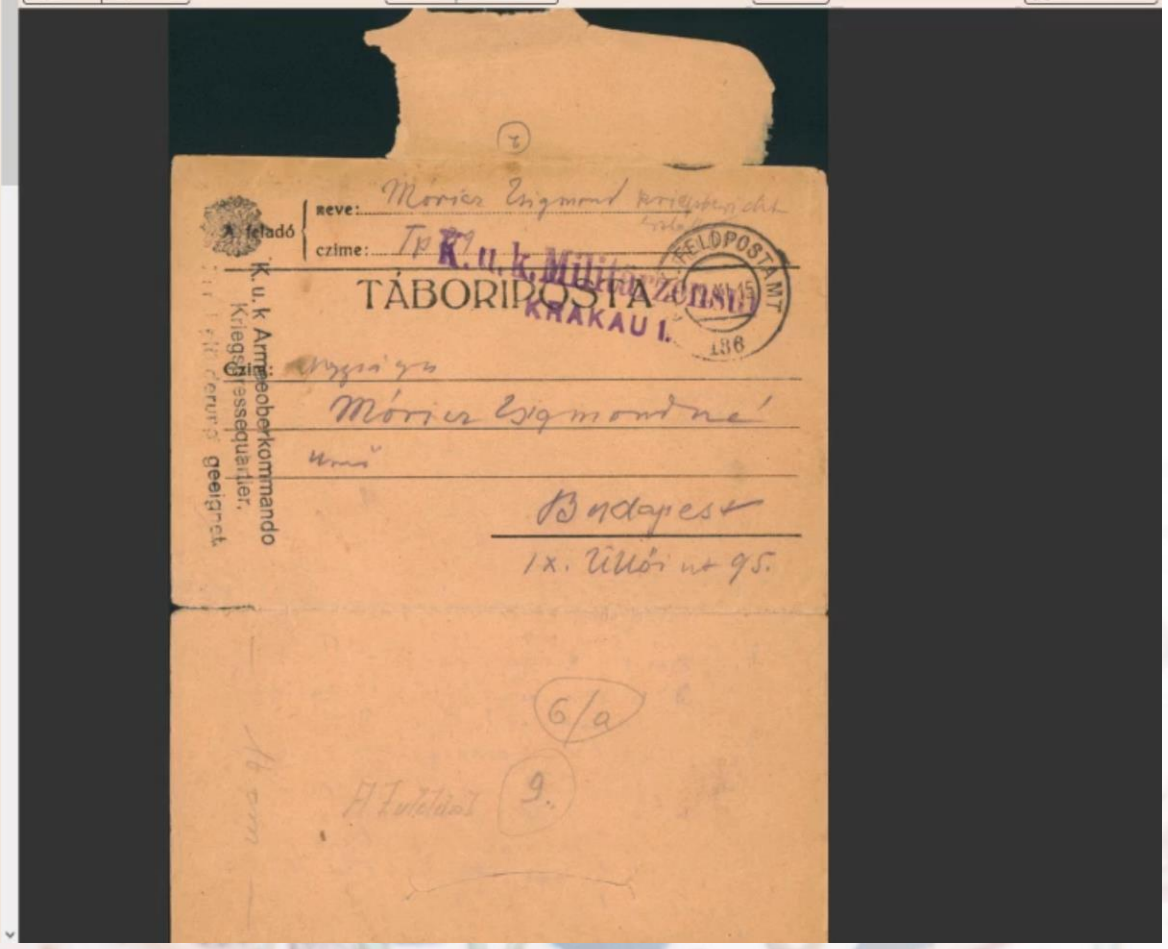
Nagyságos
Móricz Zsigmondné
urnő

Budapest
IX. Üllői út 95.*

XI. 18. II 1915.

Édes, kedves Drágaságom, ma már oly szép

Nagyít | Kicsinyít | Előző | Következő | Vissza | Teljes képernyő



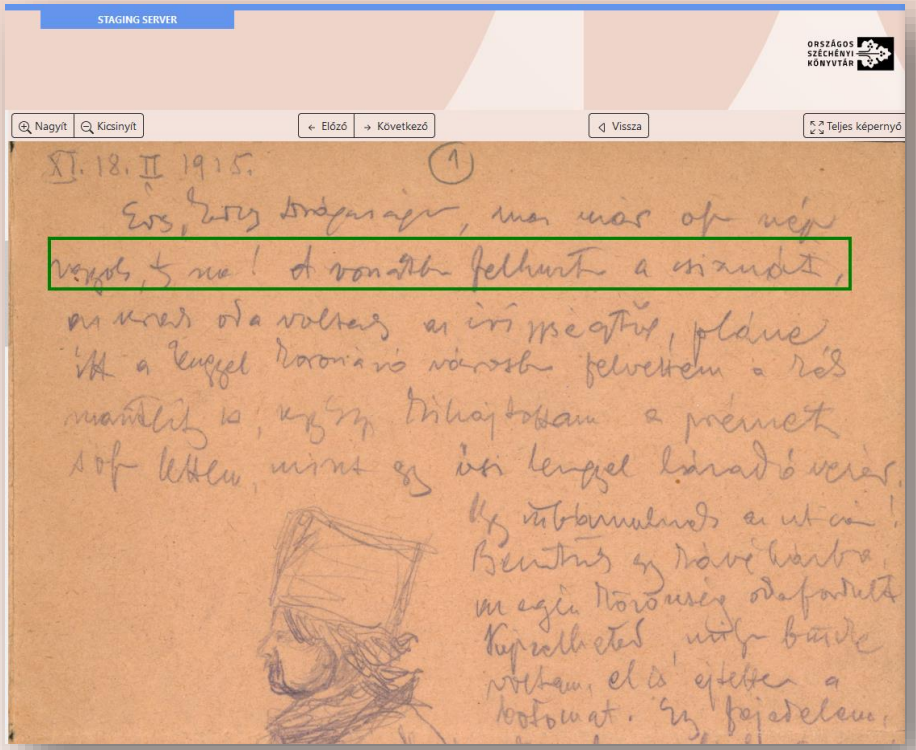
dhupla

digital humanities platform /// digitális bölcsészeti platform

XI. 18. II 1915.

Édes, kedves Drágaságom, ma már oly szép
vagyok, hogy na! A vonatban felhuztam a csizmát ,
az urak oda voltak az irigységtől, pláne
itt a lengyel koronázó városban felvettem a kék
mantlit is, ugy hogy kihajtottam a prémet
s olyan lettem mint egy úri lengyel lázadó vezér.
Ugy megbámultak az utcán!
Bementünk egy kávéházba,
az egész közönség odafordult.
Képzelheted, milyen büszke
voltam, el is ejtettem a
botomat. Egy fejedelem,
aki a bevonuláskor elejti
a jogarát és le kell
hajolnia érte. S még
csak az is kölcsön van
a Fehéri jó öreg botja.

Különben ma kaptam
először ugyane mi¬
nőségemben, mint rebellis

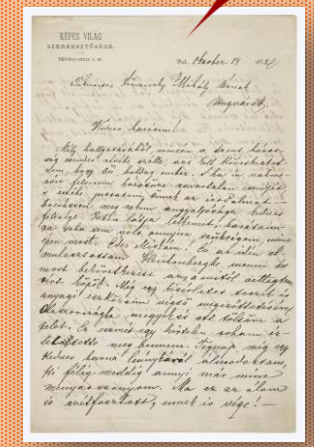


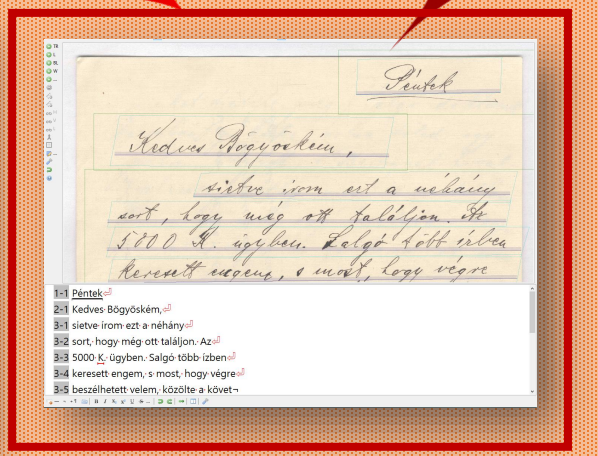

```
      tiszteletére állított emléktábla található.
      <ref target="https://resolver.pim.hu/bib/PIM1644068"/>
      A ház történetéről részletesebben lásd a Móricz-séta 11. helyszínét.
      <ref target="https://dhupla.pim.hu/page/kreativ/moricz-seta"/>
    </note>
  </p>
  <pb facs="#facs_2" n="2"/>
  <p facs="#facs_2_region_1632301340397_31">
    <lb facs="#facs_2_line_1632301452777_44" n="N001"/>
    XI. 18. II 1915.
  </p>
  <p facs="#facs_2_region_1632301390717_34">
    <lb facs="#facs_2_line_1632301459446_54" n="N001"/>
    Édes, kedves
    <persName>
      Drágaságom
      <idno corresp="Holics Janka" type="KOHA_AUTH">KOHA_AUTH:313721</idno>
    </persName>
    , ma már oly szép
    <lb facs="#facs_2_line_1632301462446_59" n="N002"/>
    vagyok,
    <seg corresp="Móricz Zsigmond" type="stenography">hogy</seg>
    na! A vonatban felhuztam a csizmát
    <ref target="https://resolver.pim.hu/bib/PIM798960"/>
    ,
    <lb facs="#facs_2_line_1632301465150_64" n="N003"/>
    az urak oda voltak az irigységtől, pláne
    <lb facs="#facs_2_line_1632301467974_69" n="N004"/>
    itt a lengyel koronázó városban felvettem a kék
    <lb facs="#facs_2_line_1632301470463_74" n="N005"/>
    mantlit is, ugy hogy kihajtottam a prémet
    <lb facs="#facs_2_line_1632301474007_79" n="N006"/>
    s olyan lettem mint egy úri lengyel lázadó vezér.
    <lb facs="#facs_2_line_1632301476864_84" n="N007"/>
    Ugy megbámultak az utcán!
    <lb facs="#facs_2_line_1632301485906_89" n="N008"/>
    Bementünk egy kávéházba,
    <lb facs="#facs_2_line_1632301488315_94" n="N009"/>
    az egész közönség odafordult.
    <lb facs="#facs_2_line_1632301490522_99" n="N010"/>
    Képzelheted, milyen büszke
    <lb facs="#facs_2_line_1632301493211_104" n="N011"/>
    voltam, el is ejtettem a
    <lb facs="#facs_2_line_1632301495444_109" n="N012"/>
    botomat. Egy fejedelem,
    <lb facs="#facs_2_line_1632301497763_114" n="N013"/>
    aki a bevonuláskor elejti
    <lb facs="#facs_2_line_1632301500131_119" n="N014"/>
    a jogarát és le kell
    <lb facs="#facs_2_line_1632301502260_124" n="N015"/>
    hajolnia érte. S még
    <lb facs="#facs_2_line_1632301504948_129" n="N016"/>
    csak az is kölcsön van
    <lb facs="#facs_2_line_1632301507278_134" n="N017"/>
    a
    <persName>
      Fehéri
      <idno corresp="Fehéri Armand" type="PIM">PIM:199864</idno>
    </persName>
    jó öreg botja.
  </p>
```

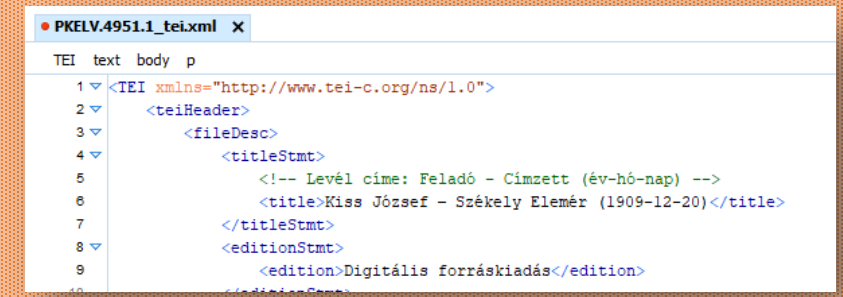
Workflow (DBK)

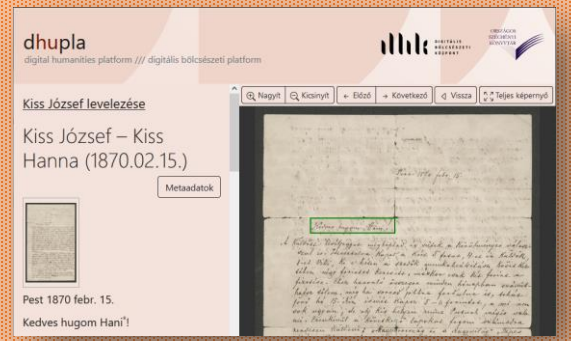
ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

dhupla









Digitalizált kézirat (szöveg → kép)

Transkribus (kép → szöveg)

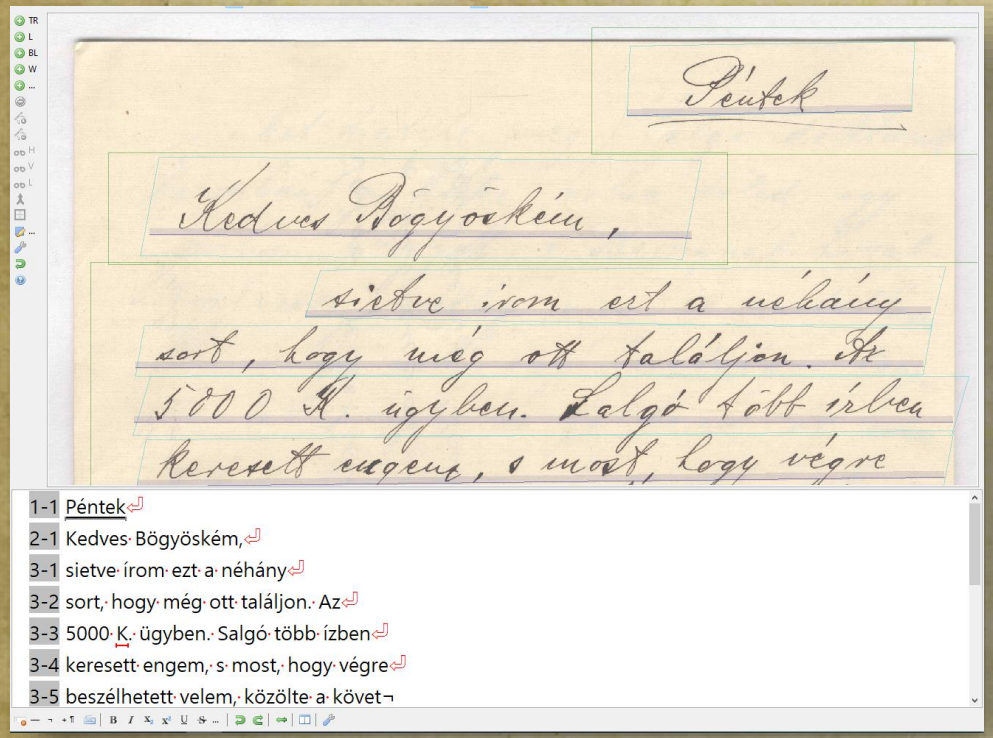
TEI XML annotált szöveg

dhupla: kép & szöveg párhuzamos megjelenítés

GitLab

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

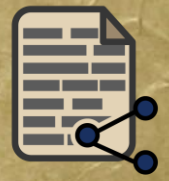
Transkribus®



Átírás

Taggelés

Együttműködés



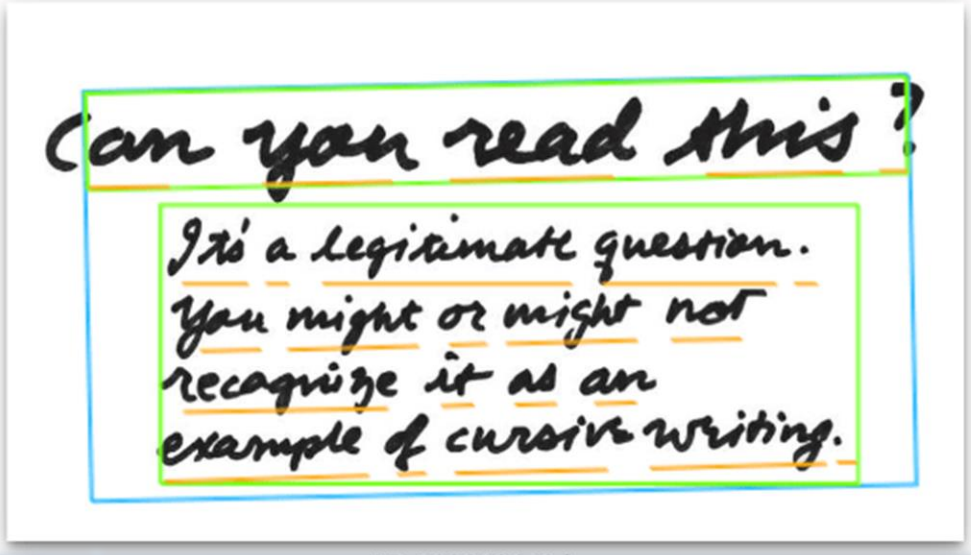
Exportálás

HANDWRITTEN TEXT RECOGNITION
WITH
Transkribus

Mi az a HTR?

Automatikus kézírásfelismertetés

ARTIFICIAL NEURAL NETWORK



AR-190319612.jpg

+Page 1

+Block 1

+Paragraph 1

can you read this ?

+Paragraph 2

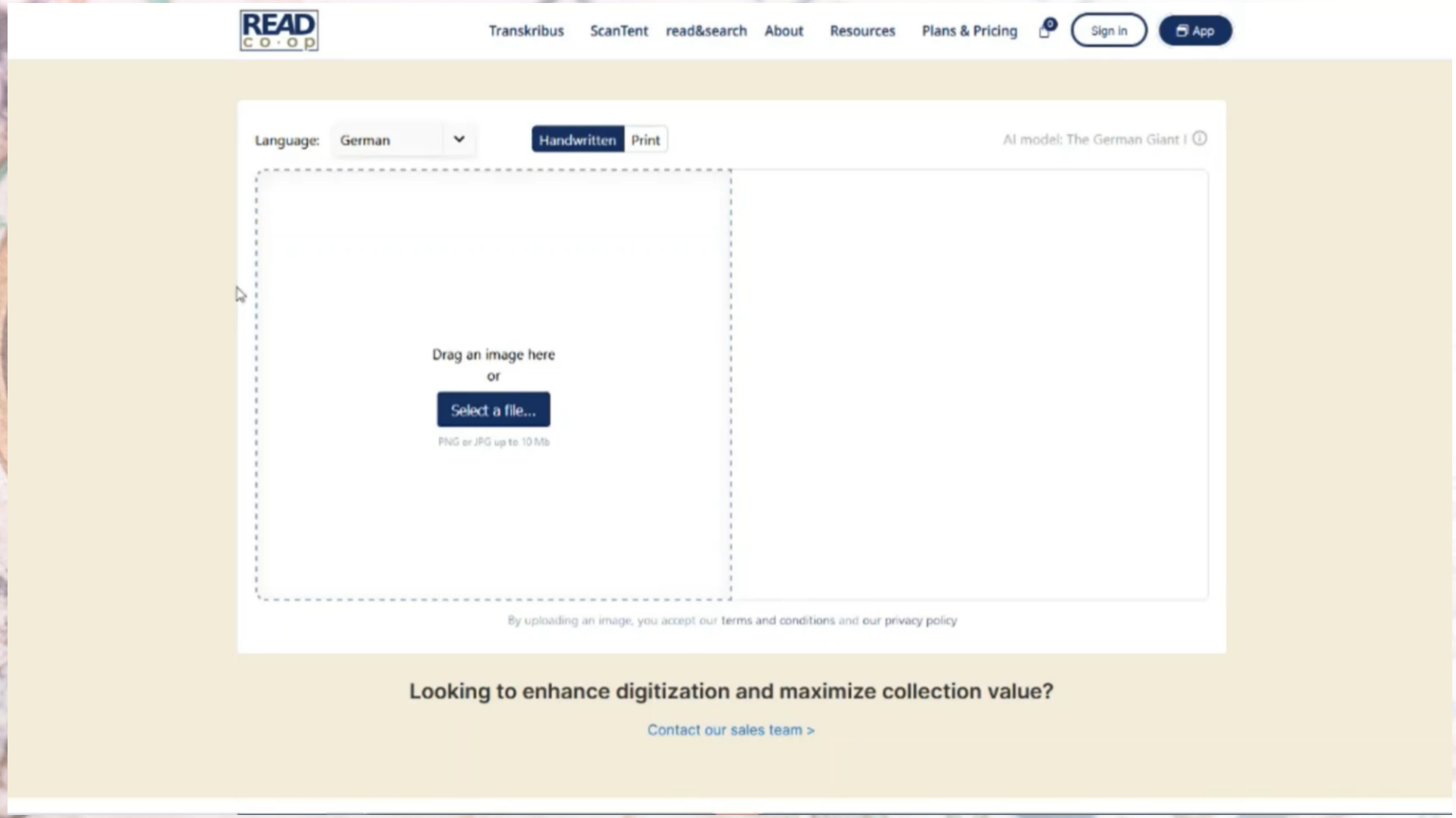
gró a legitimate questi on . you might or might not recognize it as an e xample of cursin writin g .

https://readcoop.eu/transkribus

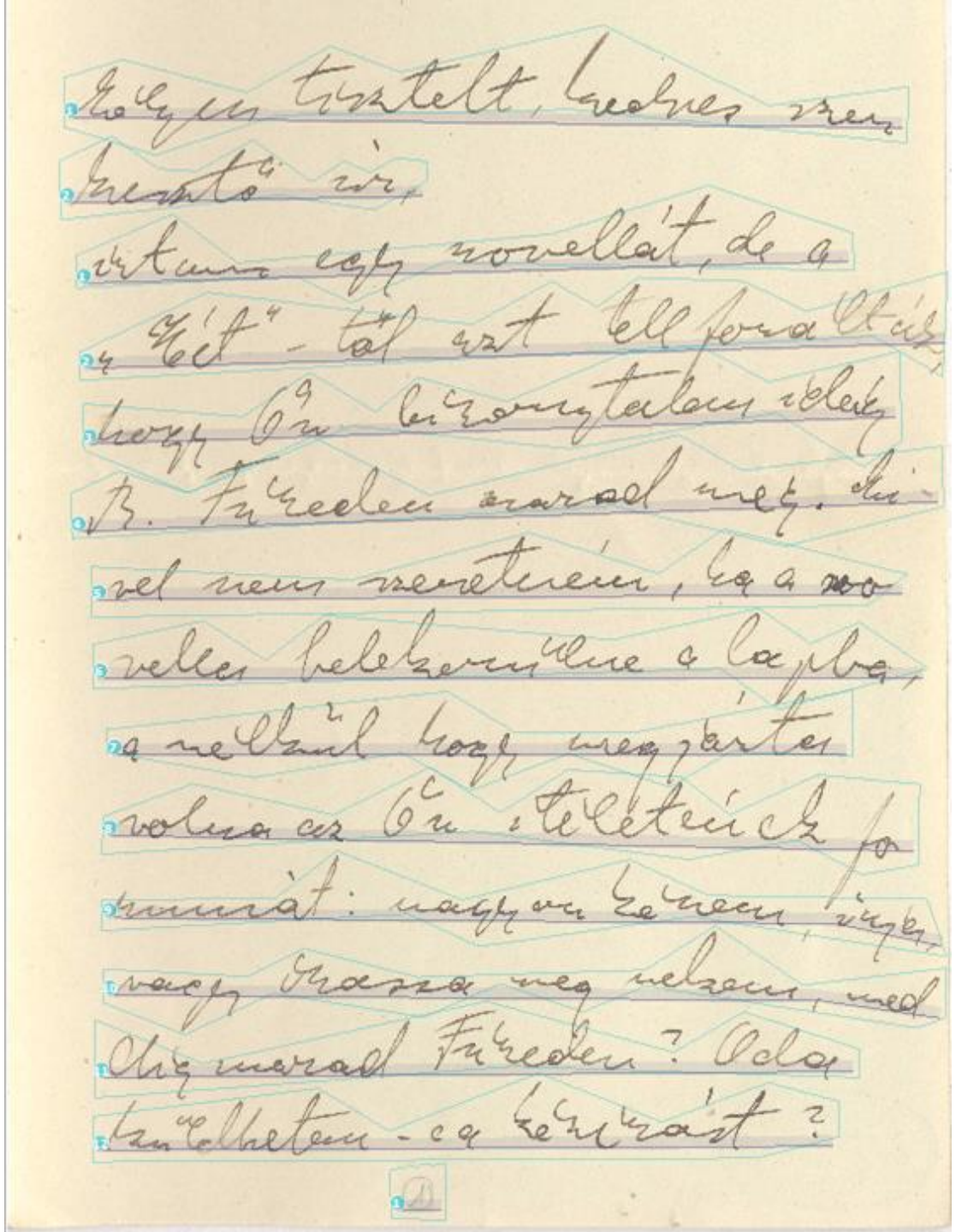
*

https://readcoop.eu/model/hungarian/

ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR



Legjobb eredmények, vegyes kézírásmodell 10%



Levélíró:
Rosenfeld Bella (1869-1908)
Címzett:
Kiss József (1843-1921)

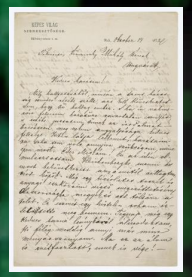
Version Comparator

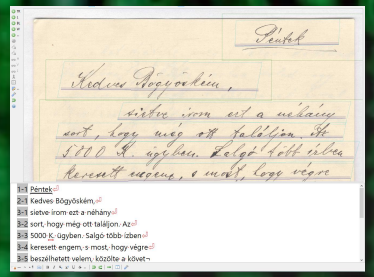
Show line numbers

```
1-1 # kélyen Mélyen tisztelt Kedves kedves szer¬
1-2 # kesztő ur úr,
2-1 # írtam egy novellát, de a
2-2 # a „Héét" töl „Hét"- től azt ellforáltúk telefonálták,
2-3 # hogy Ön bizonytalan illen Ön bizonytalan ideig
2-4 # Sz. B. Füreden sarad meg, hi¬ marad még. Mi¬
2-5 # vel nem szeretném, ha a mo¬ no¬
2-6 # velles belekenülne vella belekerülne a lapba lapba,
2-7 # a nélkül hogy megjártas megjárta
2-8 # volna az Ön íteleténeck ítéletének fo¬
2-9 # numat: rumát: nagyon kénem, vajez¬ kérem, írja,
2-10 # vagy Rassa nég nekem, írassa meg nekem, med¬
2-11 # Uig dig marad Füzeden? Pelgg Füreden? Oda
2-12 # külehetem e küldhetem-e a kénirast? kézírást?
3-1 # 1
```

V. 5239/116 https://resolver.pim.hu/bib/PIM1225955

Workflow (DBK)

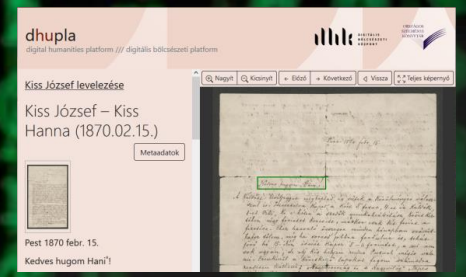




PKELV.4951.1_tei.xml

TEI text body p

```
1  <TEI xmlns="http://www.tei-c.org/ns/1.0">
2      <teiHeader>
3          <fileDesc>
4              <titleStmt>
5                  <!-- Levél címe: Feladó - Címzett (év-hó-nap) -->
6                  <title>Kiss József – Székely Elemér (1909-12-20)</title>
7              </titleStmt>
8              <editionStmt>
9                  <edition>Digitális forráskiadás</edition>
```



```
  <title>Kiss József – Tömörkény István (1908-02-05)<,t
 </titleStmt>
▼<editionStmt>
  <edition>Digitális forráskiadás</edition>
 </editionStmt>
▼<publicationStmt>
 ▼<publisher>
    <orgName>Petőfi Irodalmi Múzeum</orgName>
    <ref type="url">http://viaf.org/viaf/152132060/</ref>
    <ref type="url">http://www.pim.hu</ref>
  </publisher>
 ▼<pubPlace>
    Budapest
    <ref type="url">http://www.geonames.org/3054643</ref>
  </pubPlace>
  <date>2021</date>
 ▼<availability>
  ▼<p>
     ©In Copyright
     <ref type="url">http://rightsstatements.org/vocab/InC/1.0/</ref>
    </p>
  </availability>
  <idno type="PID">PKELV.1344_tei</idno>
  <idno type="URL">/PKELV.1344_tei.xml</idno>
 </publicationStmt>
▼<notesStmt>
 ▼<note type="critIntro" xml:id="crit.1">
  ▼<persName>
    Kiss József
    <idno corresp="Kiss József" type="PIM">PIM:61086</idno>
   </persName>
   szegedi, a Dugonics Társaságban tartandó, előadása ügyében.
  </note>
 ▼<note type="publication">
   <bibl/>
  </note>
 </notesStmt>
▼<sourceDesc>
 ▼<msDesc>
  ▼<msIdentifier>
    <country>Magyarország</country>
   ▼<settlement>
     Budapest
     <idno type="KOHA_GEO">KOHA_GEO:9227</idno>
    </settlement>
    <institution>Petőfi Irodalmi Múzeum</institution>
    <repository>Petőfi Irodalmi Múzeum Kézirattár</repository>
    <idno>V. 1344</idno>
    <msName>Szalay József-gyűjtemény</msName>
   </msIdentifier>
  ▼<msPart style="letter">
    <msIdentifier/>
   ▼<physDesc>
    ▼<objectDesc>
     ▼<supportDesc>
      ▼<extent>
```


TEI XML

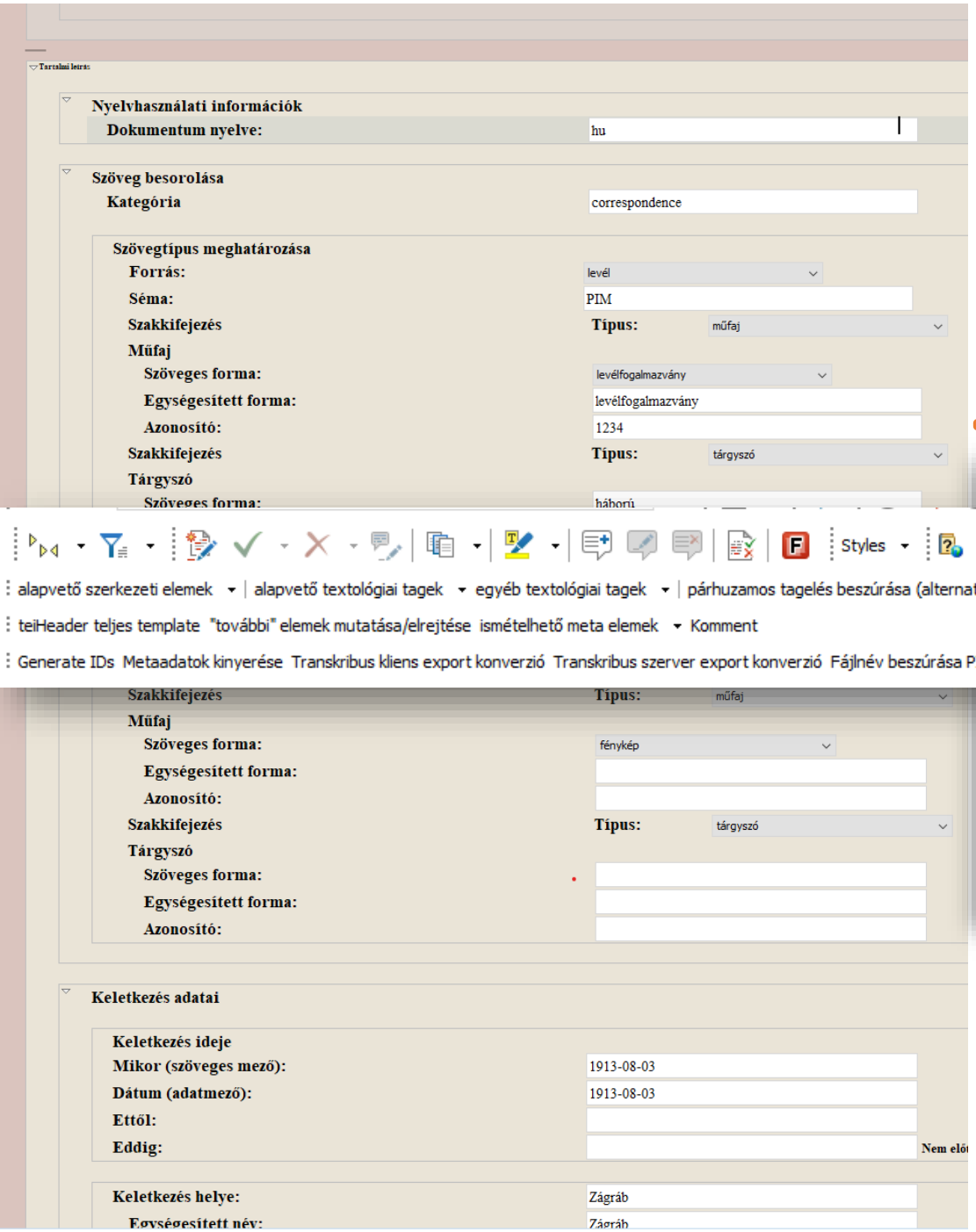
- XML: eXtensible Markup Language
- TEI: Text Encoding Initiative

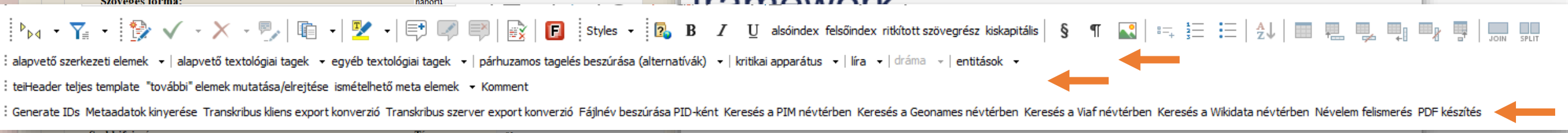
Részei

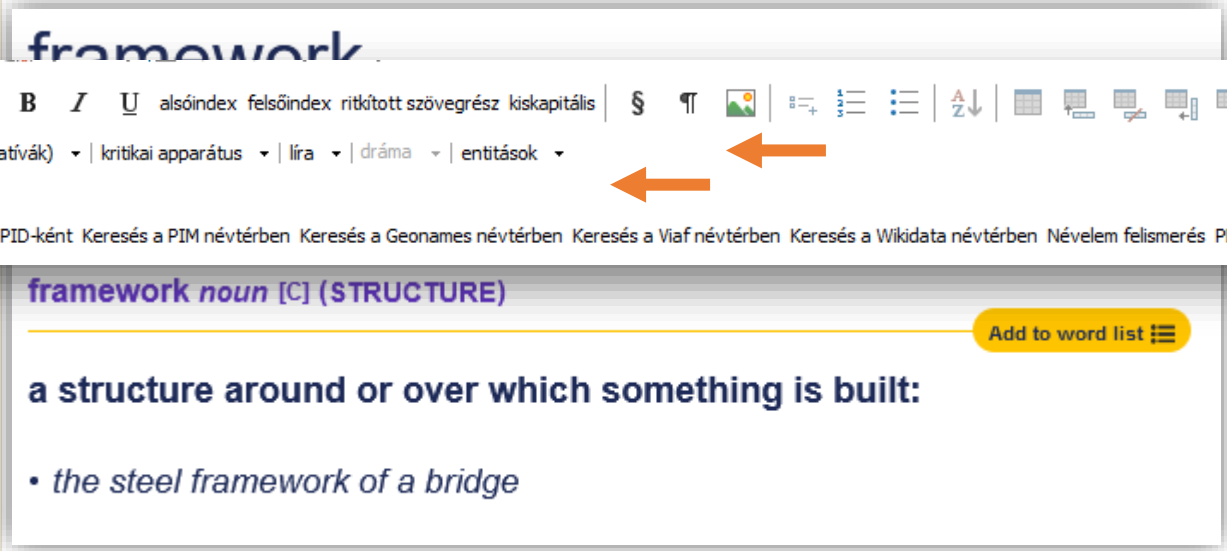
<persName>
type="PIM"

Name>

Oxygen framework

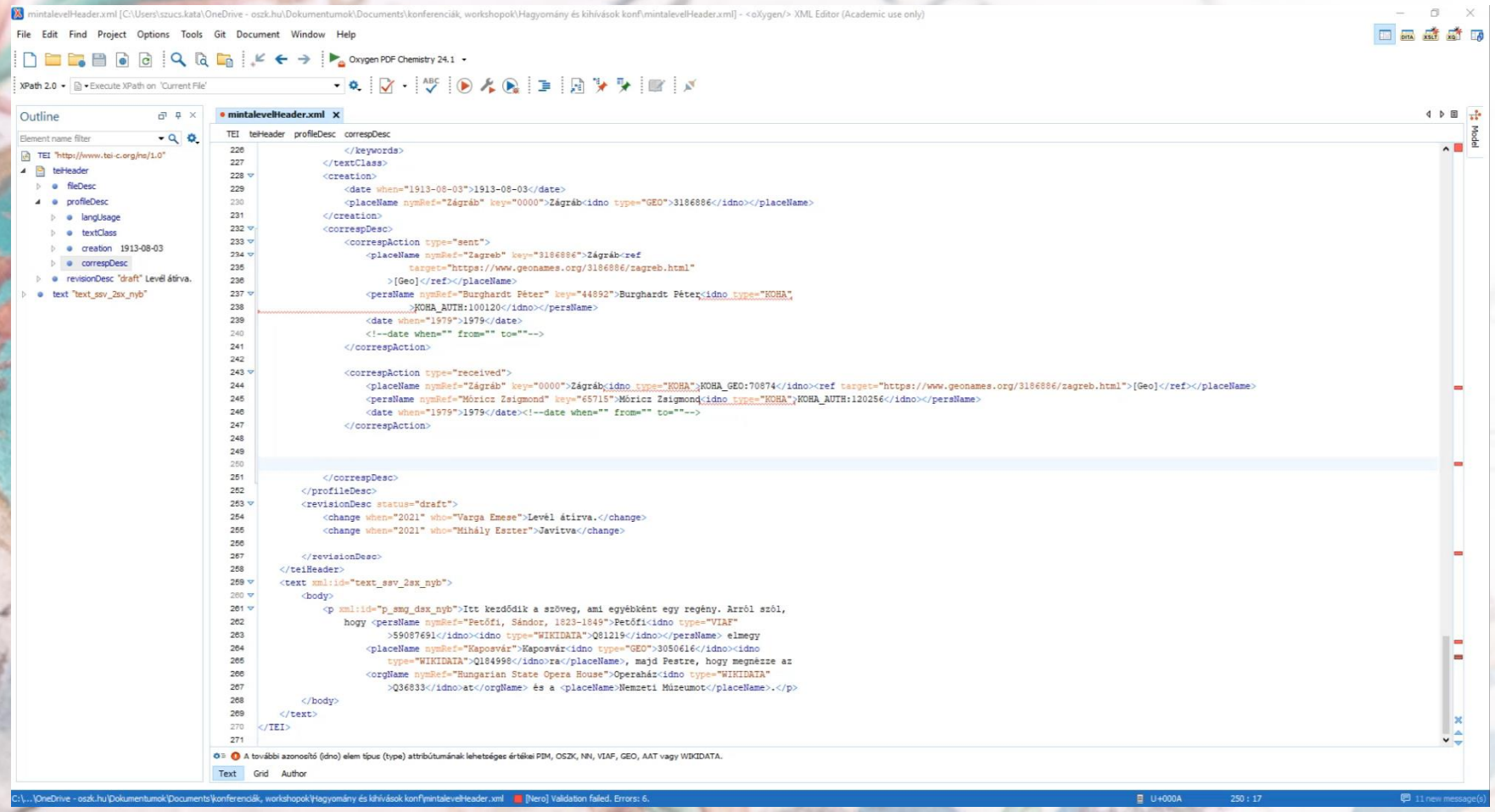






- Felhasználói környezet

Framework – segítő funkciók a szövegfeldolgozáshoz

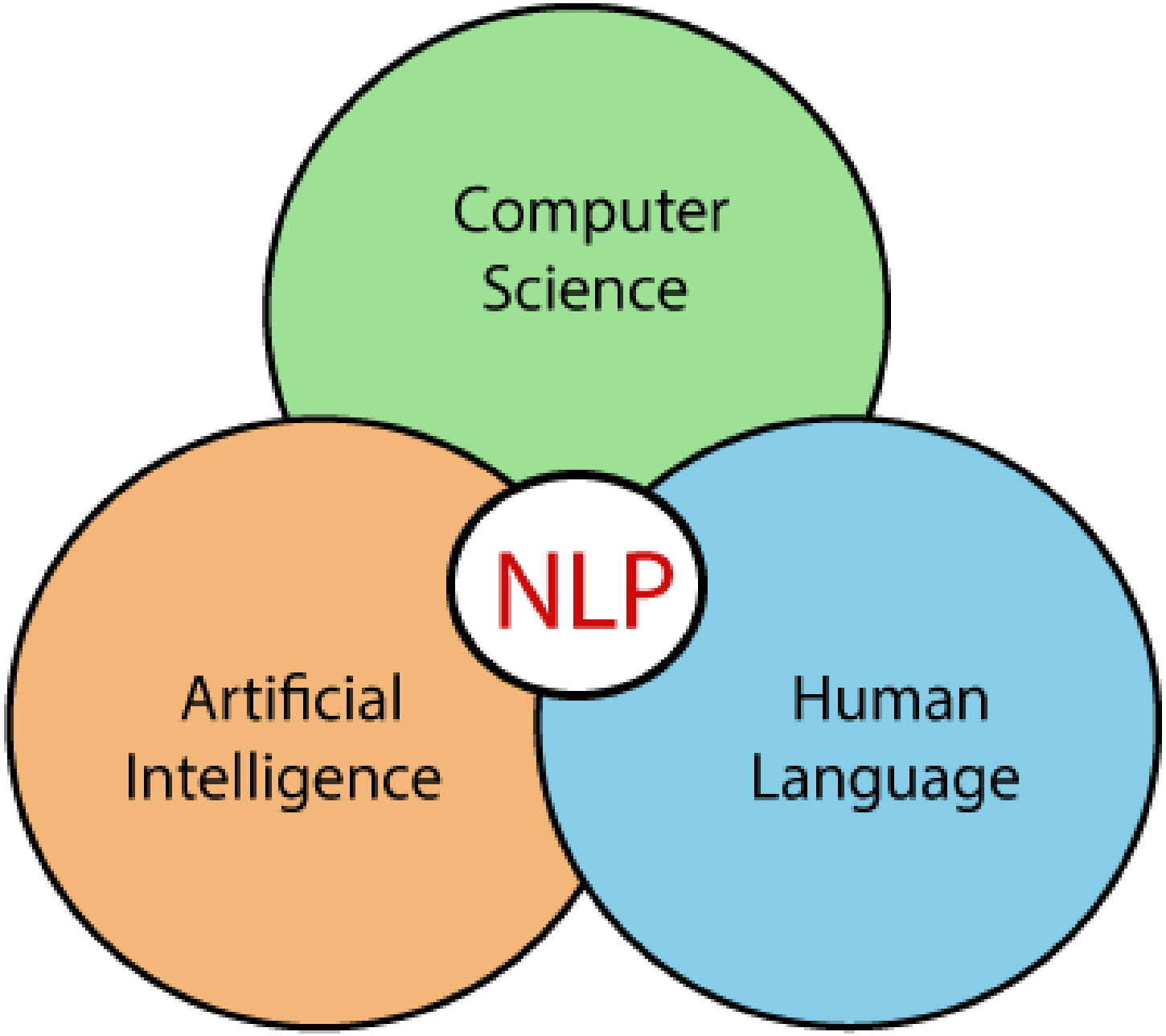


ORSZÁGOS SZÉCHÉNYI KÖNYVTÁR

Természetesnyelv-feldolgozás (Natural Language Processing, NLP)

Definíció

- a természetes nyelvű szövegek számítógépes feldolgozásával foglalkozik



Alkalmazások

A nyelvfeldolgozó eszközök tipikusan nagyobb alkalmazásokba beépítve jelennek meg:

- helyesírás-ellenőrzés: böngészők, szerkesztők
- auto-complete
- természetes nyelvű keresés a böngészőben
- gépi fordítás: Google Translate, DeepL
- automatikus beszédgenerálás a GPS alkalmazásban
- személyi asszisztensek: Siri, Alexa, Cortana, Google Assistant
- chatbotok, ChatGPT

Szövegfeldolgozási lépések

Mondatokra és szavakra bontás

- **Mondatszegmentálás:**
  - Mondathatárok azonosítása
  - Minden mondat
  - Pontos problémák, egyéb nehézségek
- **Tokenizálás:**
  - Szóalkotó karakterek és szónemalkotó karakterek
  - Számok, informatikai kifejezések, smiley-k...

https://e-magyar.hu/hu/parser

https://huggingface.co/spaces/huspacy/demo

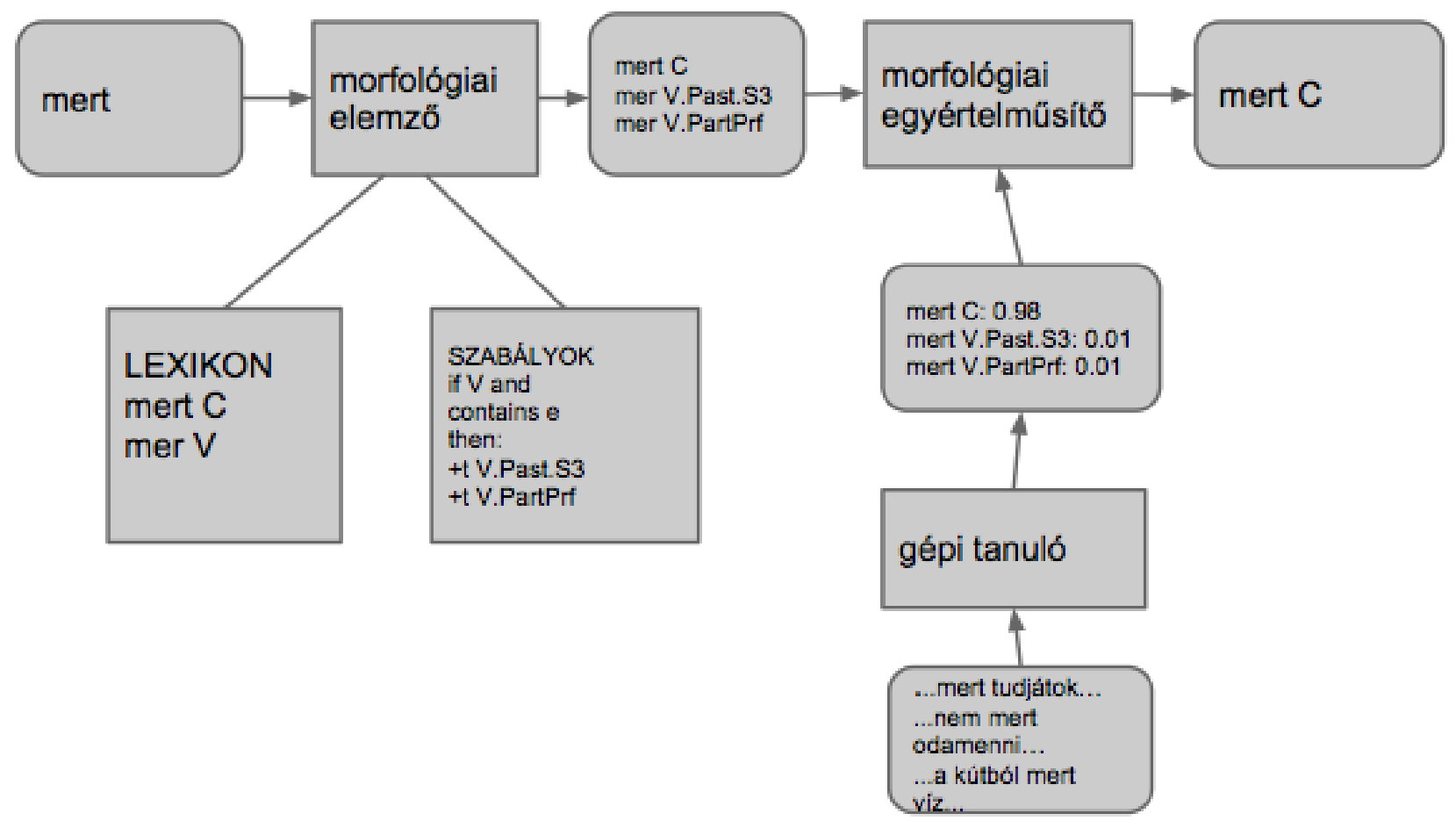
https://444.hu/2023/08/27/eloszor-futottak-maratont-2-ora-10-percen-belul-magyarorszagon-az-ugandai-victor-kiplangat-lett-a-vilagbajnok

Morfológiai elemzés

tokenszintű elemzés →nem lát se előre, se hátra →nincs kontextus →többértelműség

falucska	fa[N]luc[N]ska[N]
falucska	fa[N]lucsok[N]a[PxS3]
falucska	falucsok[N]a[PxS3]
falucska	falu[N]cska[_Dim=cskA]
falucska	falucska[N]

Morfológiai egyértelműsítés



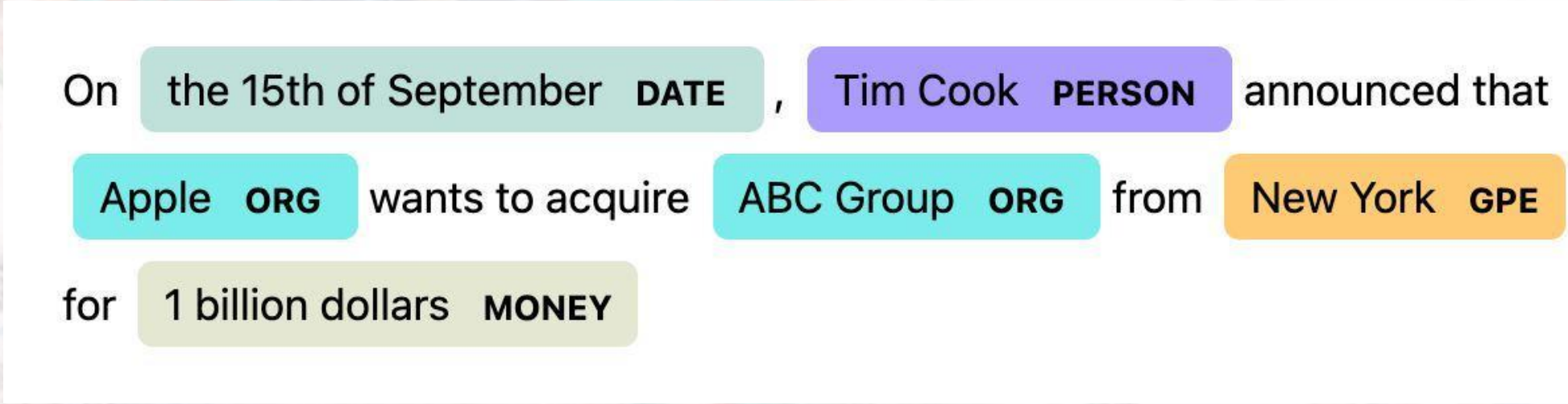
Szintaktikai elemzés

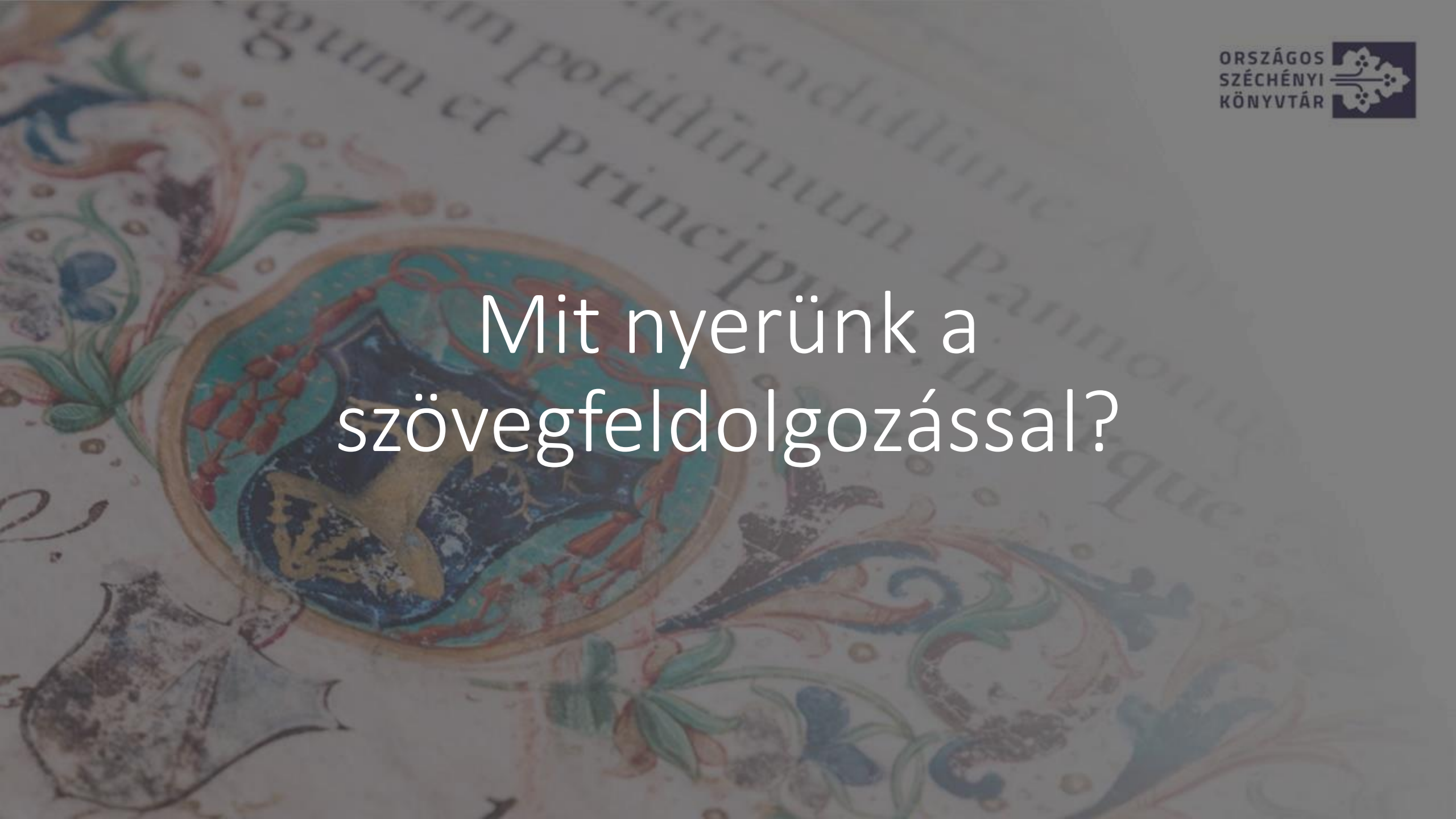
- **sekély szintaktikai elemzés:** a főnévi csoportok megtalálása
- **összetevős elemzés:** azt tárja fel, hogy a szavak egymással kombinálódva milyen kifejezéseket alkotnak, illetve hogyan állnak össze egy mondattá
- **függőségi elemzés:** a mondatok szerkezeti egységei közötti függőségi viszonyokat (pl. alany, tárgy, jelző) tárja fel

https://e-magyar.hu/hu/parser
https://huggingface.co/spaces/huspacy/demo

Tulajdonnév-felismerés (Named Entity Recognition, NER)

1. a nevek lokalizálása strukturálatlan szövegben
2. a megtalált elemek besorolása előre definiált névosztályokba

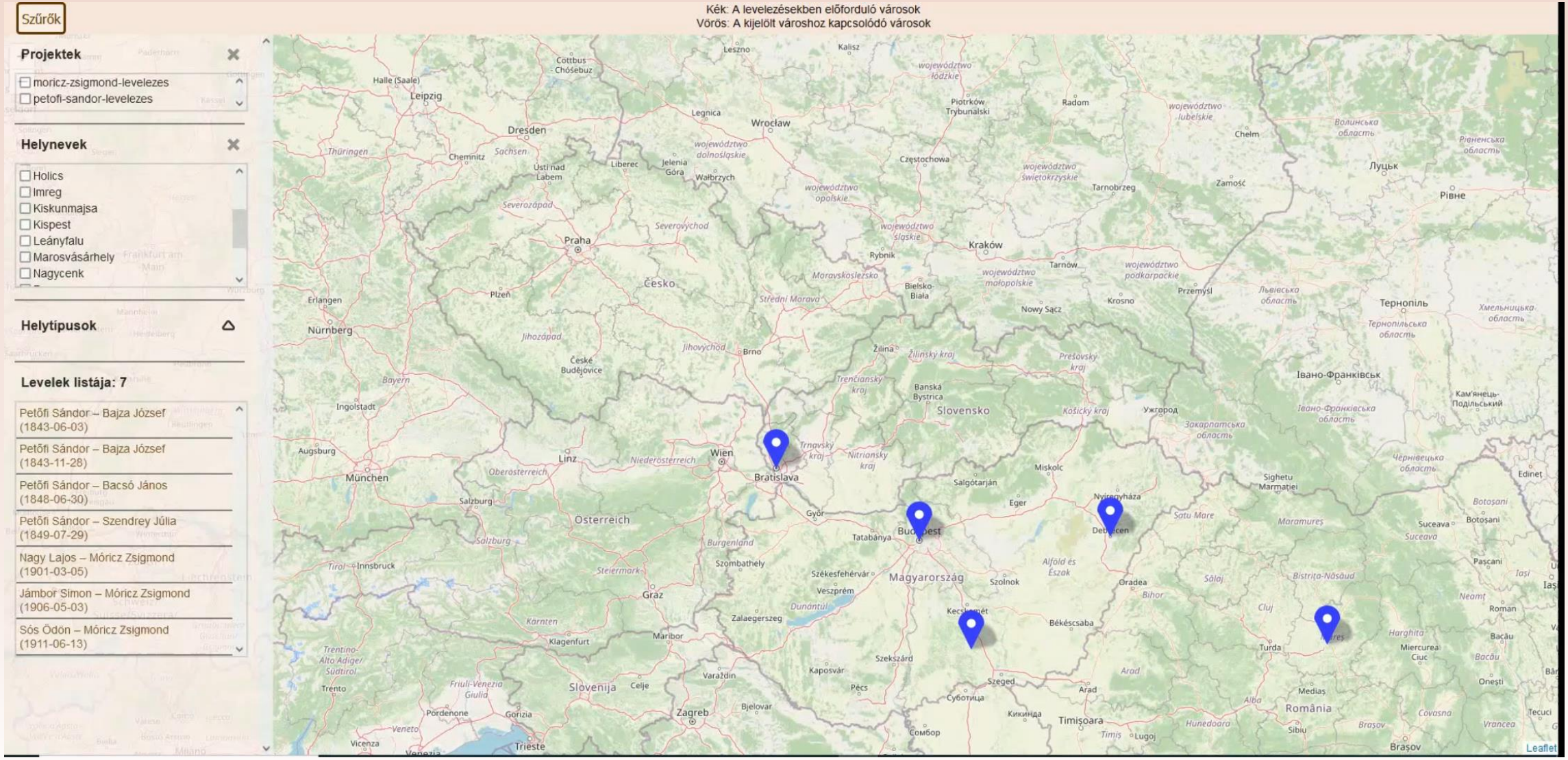




Mit nyerünk a szövegfeldolgozással?

- lesz egy mondatokra és szavakra bontott dokumentumunk:
  - szó- és mondatstatisztikák →szerzőazonosítás, stilometria
- minden szónak tudni fogjuk a tövét:
  - tőalapú keresés vs. szóalakalapú keresés →okosabb keresés, több találat
  - tőalapú statisztikák →szókincs
- minden szónak tudni fogjuk a szófaját:
  - szófajalapú statisztikák
- ismerni fogjuk a szövegben szereplő neveket:
  - megismerjük a szereplőket, helyszíneket stb.
  - térképre rakhatjuk őket
  - lehorgonyozhatjuk őket különféle adatbázisokhoz, névterekhez
- kibányászhatjuk a szövegbeli információkat, eseményeket
  - ki, hol, mikor, mit csinált?

Térképes vizualizáció



Szófelhő Kiss József levelezéséből

KJ_leveliro_gyaklista_lemma_pos.tsv

```
2929
2385 .  [/Punct]
2208 ,  [/Punct]
1339 a  [/Det|Art.Def]
567 és  [/Cnj]
470 az  [/Det|Art.Def]
458 hogy    [/Cnj]
424 nem [/Adv]
357 van [/V]
345 én  [/N|Pro]
304 !   [/Punct]
294 is  [/Adv]
268 Kiss    [/N]
248 az  [/Det|Pro]
222 József  [/N]
211 ha  [/Cnj]
195 hét [/N]
191 ön  [/N|Pro]
188 egy [/Det|Art.NDef]
182 kedves  [/Adj]
178 de  [/Cnj]
165 Budapest    [/N]
165 -
153 maga    [/N|Pro]
150 ez  [/Det|Pro]
146 lesz    [/V]
144 még [/Adv]
139 te  [/N|Pro]
125 meg [/Prev]
125 már [/Adv]
121 ő   [/N|Pro]
119 ír  [/V]
116 VIII    [/Num|Roman][_Ord/Adj]
102 :   [/Punct]
102 jó  [/Adj]
101 fog [/V]
97 csak [/Adv]
94 tud  [/V]
90 itt  [/Adv|Pro]
89 nagyon   [/Adv]
```

https://dhupla.hu/page/kreativ/kissjozsef-szofelho

Köszönjük a figyelmet!

szucs.kata@oszk.hu

simon.eszter@oszk.hu

*

https://dhupla.hu/