



dHUpla: a kéziratról 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

digital humanities platform /// digitális bölcsészeti platform

Demo

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

Metaadatok

Megjegyzések



A feladó

neve:

Móricz Zsigmond kriegsbericht-
erstatter*

címe:

Tp 39

Nagyságos

Móricz Zsigmondné
urnó

Budapest

IX. Üllői út 95.*



XI. 18. II 1915.

Édes, kedves Drágaságom, ma már oly szép

STAGING SERVER

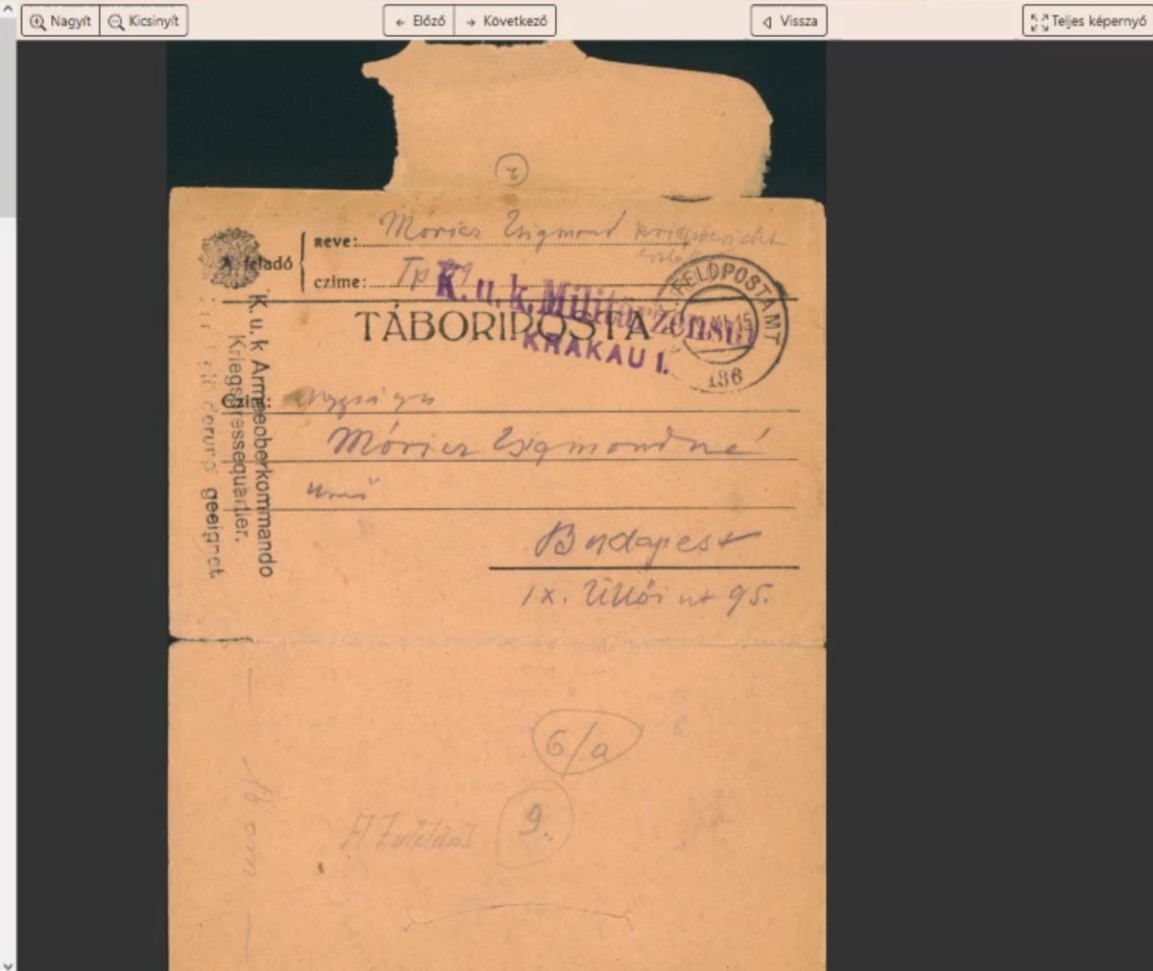
ORSZÁGOS
SZÉCHÉNYI
KÖNYVTÁR

Nagyít Kicsinyít

Előző Következő

Vissza

Tejles képernyő

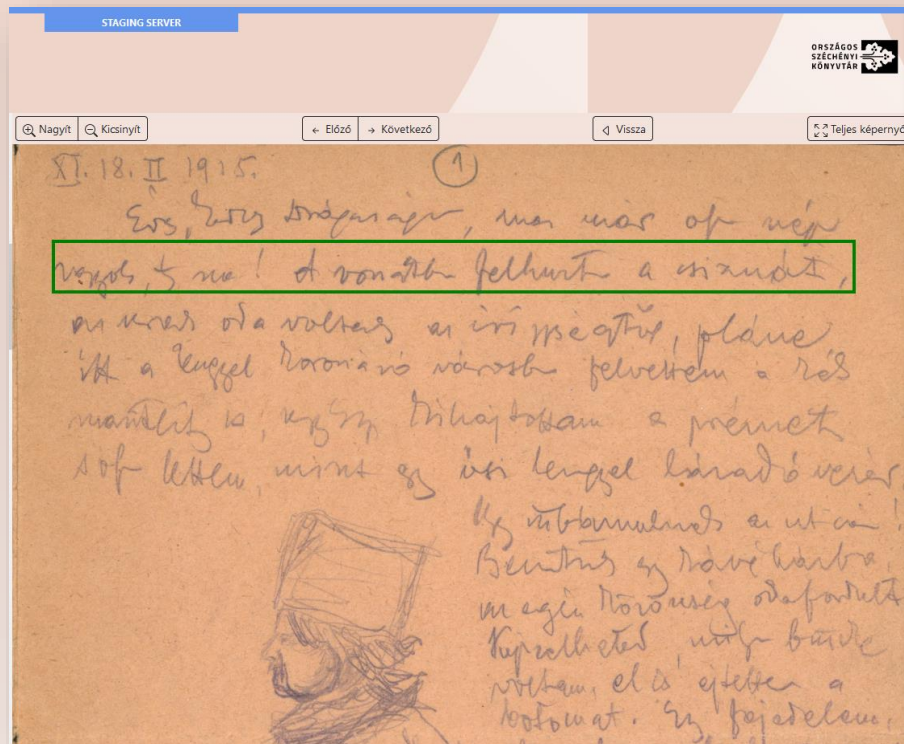




XI. 18. II 1915.

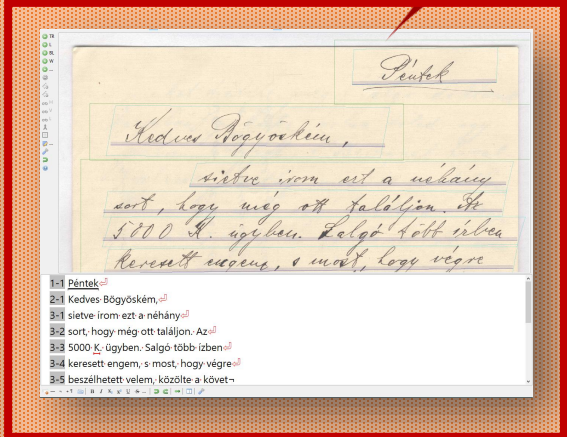
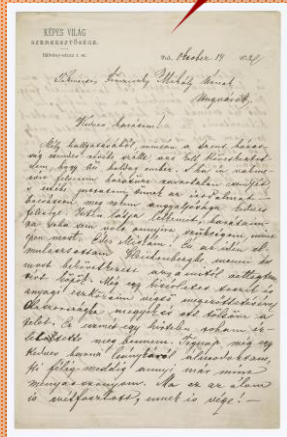
Édes, kedves Drágaságom, ma már oly szép vagyok, hogy na! A vonatban felhúztam a **csizmát**, az urak oda voltak az irigységtől, pláne itt a lengyel koronázó városban felvettem a kék mantlit is, úgy hogy kihajtottam a prémet s olyan lettem mint egy úri lengyel lázadó vezér. Ugy megbámultak az utcán! 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



```
</persName>
tiszeletére állított emléktábla található.
<ref target="https://resolver.pim.hu/bib/PIM1644068"/>
A ház történetéről részletesebben lásd a Möricz-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="Möricz Zsigmond" type="stenography">hogy</seg>
na! A vonatban felhúztam 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, úgy 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)



```
PKELV.4951.1_tei.xml X
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>
10 </editionStmt>
```

dhupla



Digitalizált kéziratos (szöveg → kép)

Transkribus (kép → szöveg)

TEI XML annotált szöveg

dhupla: kép & szöveg párhuzamos megjelenítés





Transkribus[®]

The screenshot displays a digital workspace for a handwritten document. At the top right, the word "Péntek" is written in cursive and enclosed in a blue box. Below it, several lines of cursive text are shown, each with a blue rectangular selection box. The text includes: "Kedves Bögyöském,", "sietve írom ezt a néhány", "sort, hogy még ott találjon. Az", "5000 K. ügyben. Salgó több ízben", and "keresett engem, s most, hogy végre". On the left side, there is a vertical toolbar with various icons for editing and navigation. At the bottom, a list of annotations is visible, showing line numbers and the corresponding text segments.

1-1 Péntek
2-1 Kedves Bögyöském,
3-1 sietve írom ezt a néhány
3-2 sort, hogy még ott találjon. Az
3-3 5000 K. ügyben. Salgó több ízben
3-4 keresett engem, s most, hogy végre
3-5 beszélhettem velem, közölte a követ



Átírás



Taggelés



Együttműködés



Exportálás

HANDWRITTEN TEXT RECOGNITION

WITH

Transkribus

Mi az a HTR?

Automatikus kézírásfelismertetés

ARTIFICIAL NEURAL NETWORK

Can you read this?

It's a legitimate question.
You might or might not
recognize it as an
example of cursive writing.

+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/>



Language: German AI model: The German Giant |

Drag an image here
or

PNG or JPG up to 10 Mb

By uploading an image, you accept our [terms and conditions](#) and our [privacy policy](#)

Looking to enhance digitization and maximize collection value?

[Contact our sales team >](#)

Kélyen tisztelt, kedves szer-
kesztő úr,
irtam egy novellát, de a
„Hét”-től azt elfordították,
hogy Ön bizonytalan ideig
B. Füreden marad még, de
vel nem szeretném, ha a no-
velles belekerülne a lapba,
a nélkül hogy megjárta
volna az Ön ítéletének fo-
rumát: nagyon kérem, írja,
vagy írassa meg nekem, med-
dig marad Füreden? Oda
küldhetem-e a kézirat?

Levélíró:

Rosenfeld Bella (1869-1908)

Címzett:

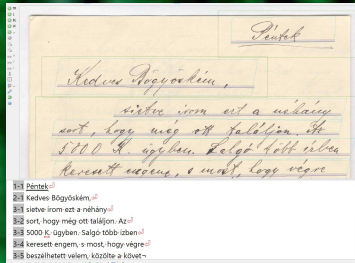
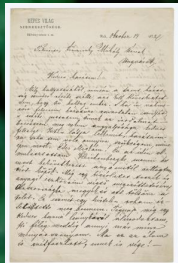
Kiss József (1843-1921)

Version Comparator

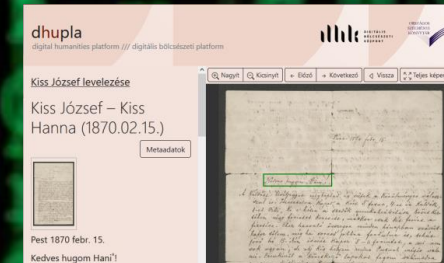
Show line numbers

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

Workflow (DBK)



```
PKELV.4951.1_tei.xml x
TEI text body p
1 <TEI xmlns="http://www.tei-c.org/ns/1.0">
2 <teiHeader>
3 <fileDesc>
4 <titleStm>
5 <!-- Levél címe: Feladó - Címzett (év-hó-nap) -->
6 <title>Kiss József - Székely Elemér (1909-12-20)</title>
7 </titleStm>
8 <editionStm>
9 <edition>Digitális forráskiadás</edition>
```



```

<title>Kiss József - Tömörkény István (1908-02-05)</title>
</title>
<edition>Digitális forráskiadás</edition>
</edition>
<publication>
  <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>
</publication>
<notes>
  <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>
</notes>
<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

Tartalmi leírás

Nyelvhasználati információk
Dokumentum nyelve: hu

Szöveg besorolása
Kategória: correspondence

Szövegtípus meghatározása

Forrás: levél

Séma: PIM

Szakkifejezés
Típus: műfaj

Műfaj
Szöveges forma: levélfogalmazvány

Egységesített forma: levélfogalmazvány

Azonosító: 1234

Szakkifejezés
Típus: tárgyszó

Tárgyszó
Szöveges forma: háború

Navigation and toolbars:

- Navigation icons: back, forward, search, etc.
- Text formatting: Bold (B), Italic (I), Underline (U), text color, background color, text alignment, bulleted list, numbered list, link, unlink, table, table of contents, etc.
- Dropdown menu: alapvető szerkezeti elemek, alapvető textológiai tagok, egyéb textológiai tagok, párhuzamos tagelés beszúrása (alternatívák), kritikai apparátus, líra, dráma, entitások
- Footer: Generate IDs, Metaadatok kinyerése, Transkribus kliens export konverzió, Transkribus szerver export konverzió, Fájlnév beszúrása PID-ként, Keresés a PIM névtérben, Keresés a Geonames névtérben, Keresés a Vial névtérben, Keresés a Wikidata névtérben, Névelem felismerés, PDF készítés

Szakkifejezés
Típus: műfaj

Műfaj
Szöveges forma: fénykép

Egységesített forma:

Azonosító:

Szakkifejezés
Típus: tárgyszó

Tárgyszó
Szöveges forma:

Egységesített forma:

Azonosító:

Keletkezés adatai

Keletkezés ideje

Mikor (szöveges mező): 1913-08-03

Dátum (adatmező): 1913-08-03

Ettől:

Eddig: Nem elő

Keletkezés helye: Zágráb

Egységesített név: Zágráb

framework *noun* [C] (STRUCTURE)

Add to word list

a structure around or over which something is built:

- *the steel framework of a bridge*

- Felhasználói környezet

Framework – segítő funkciók a szövegfeldolgozáshoz

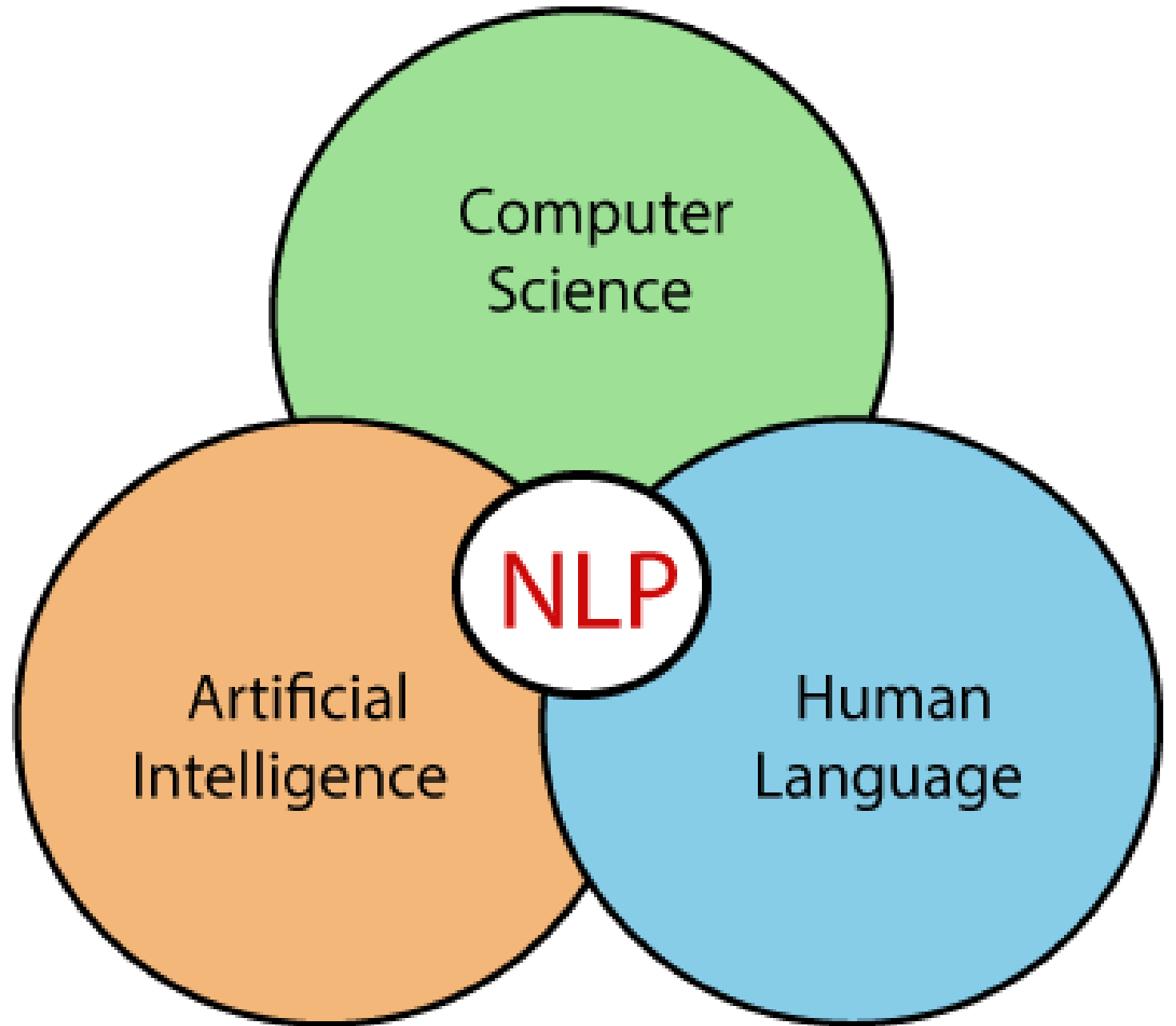
The screenshot displays the XML Editor interface for the file 'mintalevelHeader.xml'. The main window shows the XML code for the TEI header, including elements like <keywords>, <textClass>, <creation>, <correspDesc>, <correspAction>, <revisionDesc>, and <text>. The code includes references to external resources like geonames.org and Wikidata. The left sidebar shows the Outline view with a tree structure of the document elements. The bottom status bar indicates '11 new message(s)'.

```
TEI teiheader profileDesc correspDesc
226 </keywords>
227 </textClass>
228 <creation>
229 <date when="1913-08-03">1913-08-03</date>
230 <placeName nymRef="Zágráb" key="0000">Zágráb<idno type="GEO">3186886</idno></placeName>
231 </creation>
232 <correspDesc>
233 <correspAction type="sent">
234 <placeName nymRef="Zágráb" key="3186886">Zágráb<ref
235 target="https://www.geonames.org/3186886/zagreb.html"
236 [Geo]</ref></placeName>
237 <persName nymRef="Burghardt Péter" key="44892">Burghardt Péter<idno type="KOHA"
238 KOHA_AUTH:100120</idno></persName>
239 <date when="1979">1979</date>
240 <!--date when="" from="" to=""-->
241 </correspAction>
242
243 <correspAction type="received">
244 <placeName nymRef="Zágráb" key="0000">Zágráb<idno type="KOHA">KOHA_GEO:70874</idno><ref target="https://www.geonames.org/3186886/zagreb.html">[Geo]</ref></placeName>
245 <persName nymRef="Móricz Zsigmond" key="65715">Móricz Zsigmond<idno type="KOHA">KOHA_AUTH:120256</idno></persName>
246 <date when="1979">1979</date><!--date when="" from="" to=""-->
247 </correspAction>
248
249
250
251 </correspDesc>
252 </profileDesc>
253 <revisionDesc status="draft">
254 <change when="2021" who="Varga Emese">Lével átírva.</change>
255 <change when="2021" who="Mihály Ester">Javitva.</change>
256 </revisionDesc>
257 </teiheader>
258 <text xml:id="text_ssv_2sx_nyb">
259 <body>
260
261 <p xml:id="p_smg_dsx_nyb">Itt kezdődik a szöveg, ami egyébként egy regény. Arról szól,
262 hogy <persName nymRef="Petőfi, Sándor, 1823-1849">Petőfi<idno type="VIAF"
263 >59087691</idno><idno type="WIKIDATA">Q81219</idno></persName> elmegy
264 <placeName nymRef="Kaposvár">Kaposvár<idno type="GEO">3050616</idno><idno
265 type="WIKIDATA">Q184998</idno></placeName>, majd Pestre, hogy megnézze az
266 <orgName nymRef="Hungarian State Opera House">Operaház<idno type="WIKIDATA"
267 >Q36833</idno>at</orgName> és a <placeName>Nemzeti Múzeumot</placeName>.</p>
268 </body>
269 </text>
270 </TEI>
271
```

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>

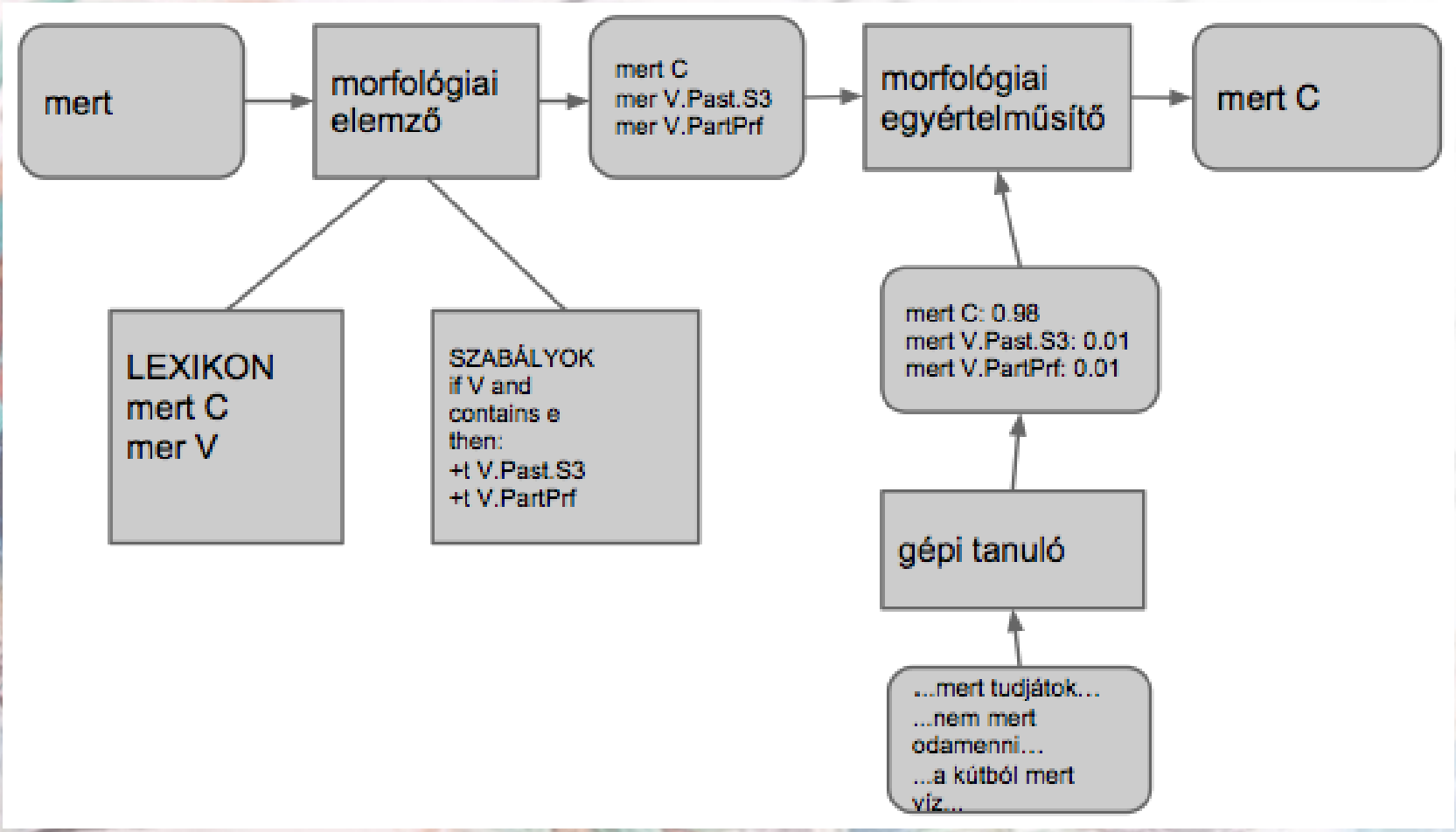


Morfológiai elemzés

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

fa1ucska	fa[N]1uc[N]ska[N]
fa1ucska	fa[N]1ucsok[N]a[PxS3]
fa1ucska	fa1ucsok[N]a[PxS3]
fa1ucska	fa1u[N]cska[_Dim=cskA]
fa1ucska	fa1ucska[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

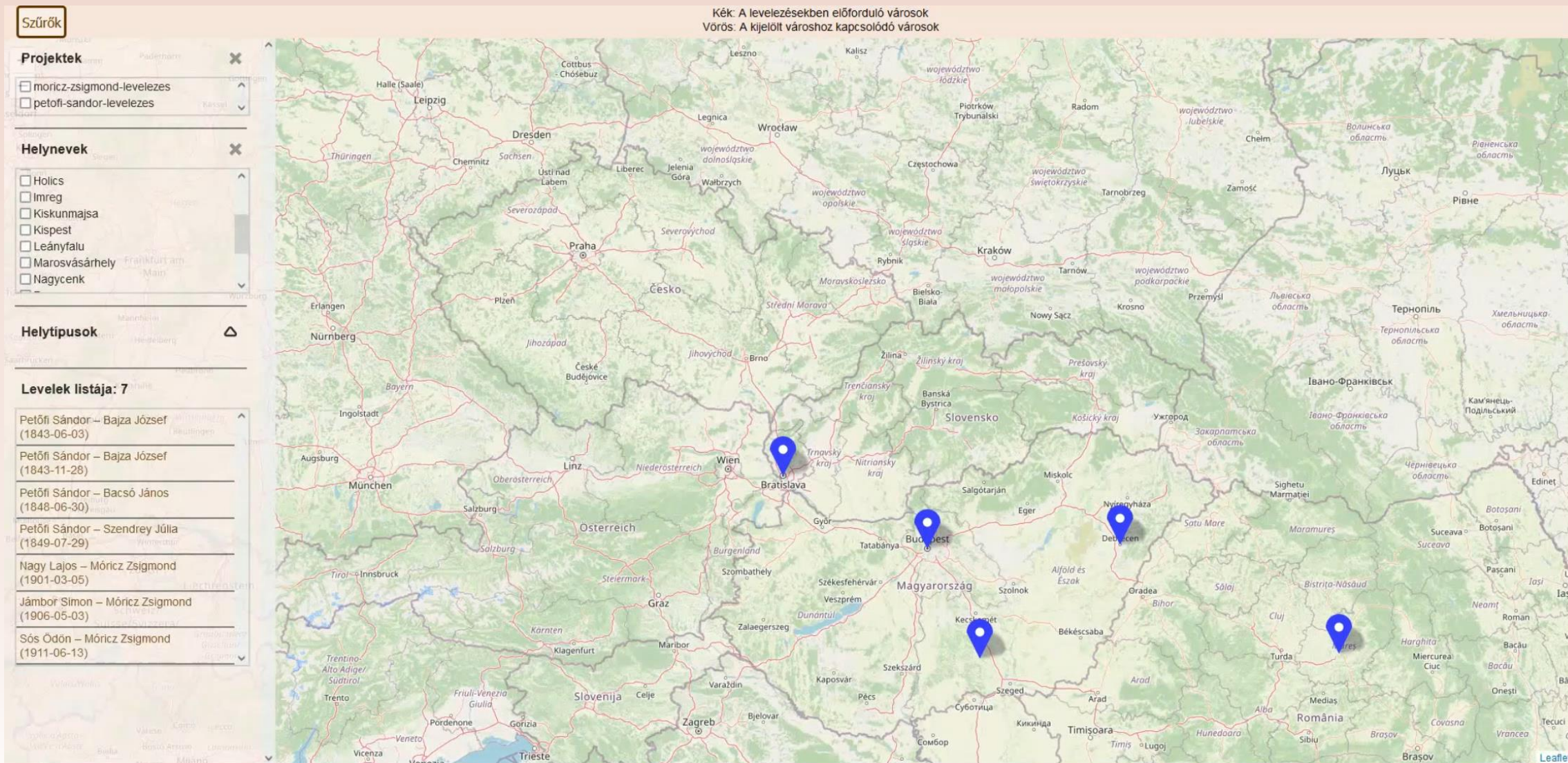
On **the 15th of September** **DATE** , **Tim Cook** **PERSON** announced that **Apple** **ORG** wants to acquire **ABC Group** **ORG** from **New York** **GPE** for **1 billion dollars** **MONEY**



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





Köszönjük a figyelmet!

szucs.kata@oszk.hu

simon.eszter@oszk.hu

*

<https://dhupla.hu/>