European Literary Text Collection (ELTeC)

DISTANT READING FOR EUROPEAN LANGUAGES

Bence Vétek, ELTE.DH

Content

- Cost Action Distant Reading Project
- Introduction of ELTeC
- The Hungarian Collection
- XML and TEI XML

COST ACTION

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

- European Cooperation in Science and Technology (1971-)
- Research network between European partners
- Interdisciplinar and open research teams
- It funds workshops, conferences, working group meetings (e.g. Third Distant Reading Training School 2019, ELTE.DH, Budapest)

COST ACTION

cost
EUROPEAN COOPERATION
IN SCIENCE & TECHNOLOGY

DISTANT READING FOR EUROPEAN LITERARY HISTORY

Distant Reading

- Aim: developing resources and methods necessary to change the way European literary history is written
- Distant Reading: using computational methods of analysis for large collections of literary texts
- Creation of a broader, more inclusive and better-grounded account of European literary history and cultural identity
- Also develops the way institutions make their holdings available to researchers

DISTANT READING FOR EUROPEAN LITERARY HISTORY

Distant Reading

The Action will:

1. build a multilingual European Literary Text Collection (ELTeC)
   - around 2,500 full-text novels
   - In at least 10 different languages,
   - permitting to test methods and compare results across national traditions;
2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe’s multilingual literary traditions;
3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

ELTeC – Sampling Criteria (Corpus Design)

Principles

- Maximize the variety
- Prefer novels published as a book over in serial publications
- No translations
- Only freely available texts, trying to reuse already digitized ones
- Non-normative sampling criteria: both canonical (30%) and non-canonical novels (30%)

ELTeC – Sampling Criteria (Corpus Design)

- Date:
  - T1: 1840-1859
  - T2: 1860-1879
  - T3: 1880-1899
  - T4: 1900-1920
- Author gender:
  - Male (M)
  - Female (F)
  - Undefined or more than one author (mixed)

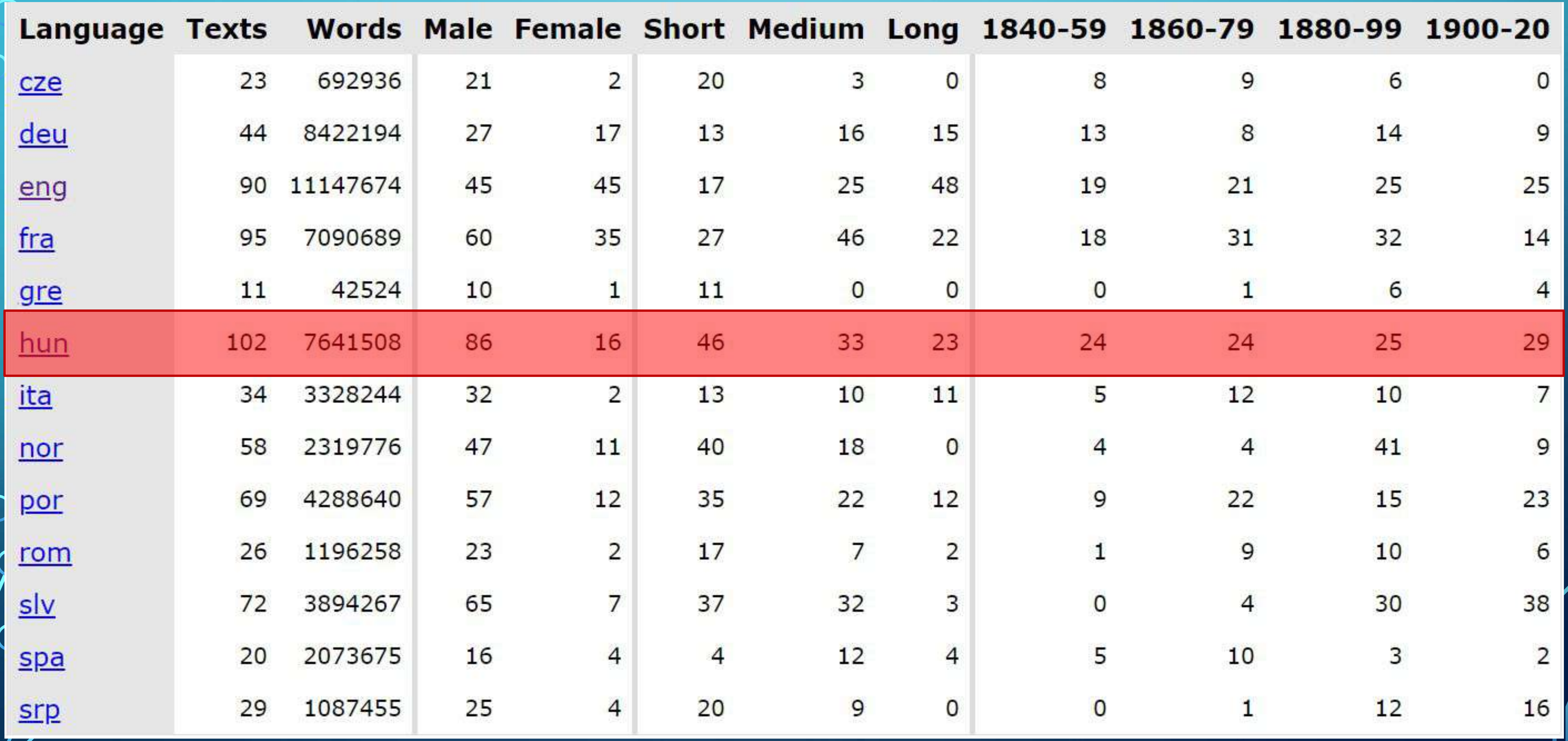
ELTeC – Sampling Criteria (Corpus Design)

- Length:
  - Short: 10k-50k word tokens
  - Medium: 50k-100k word tokens
  - Long: over 100k word tokens
- Reprint count:
  - Low: no reprints at all
  - Medium: reprinted once
  - High: reprinted more than once

ELTeC – Language

- No distinction between regional or geographical variation (e.g. Swiss German goes to the German collection)
- Only European variations (excludes US English or Quebecois)
- Language-based approach

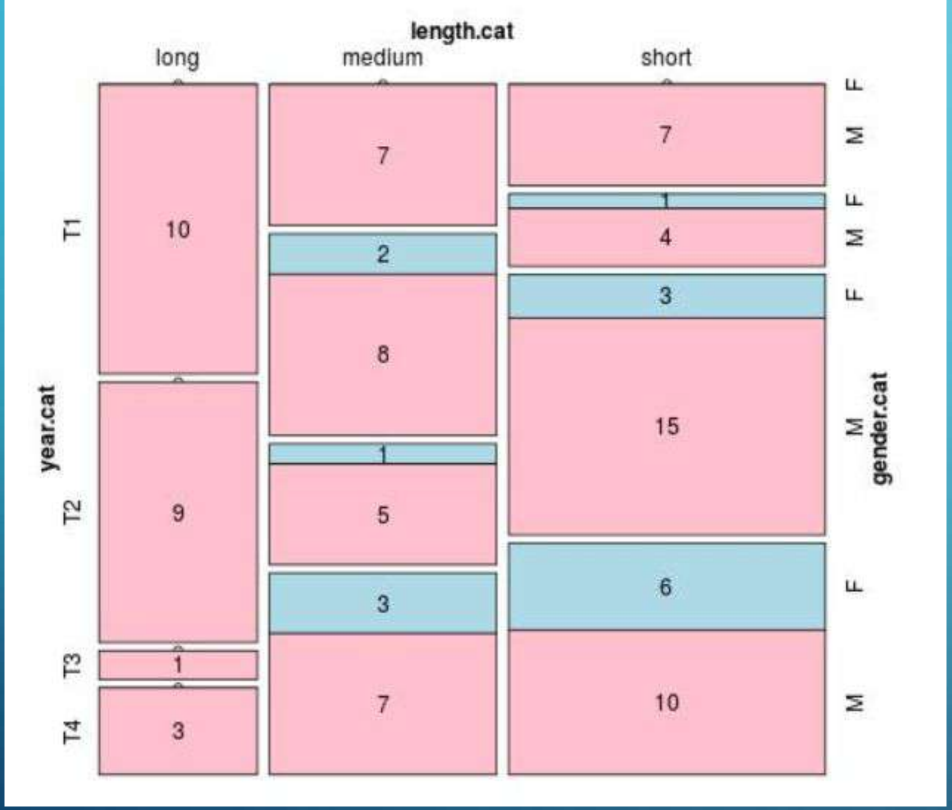
ELTeC – How does it look like?



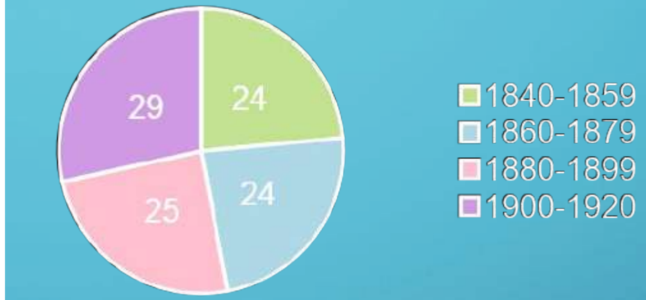
Language	Texts	Words	Male	Female	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20
cze	23	692936	21	2	20	3	0	8	9	6	0
deu	44	8422194	27	17	13	16	15	13	8	14	9
eng	90	11147674	45	45	17	25	48	19	21	25	25
fra	95	7090689	60	35	27	46	22	18	31	32	14
gre	11	42524	10	1	11	0	0	0	1	6	4
hun	102	7641508	86	16	46	33	23	24	24	25	29
ita	34	3328244	32	2	13	10	11	5	12	10	7
nor	58	2319776	47	11	40	18	0	4	4	41	9
por	69	4288640	57	12	35	22	12	9	22	15	23
rom	26	1196258	23	2	17	7	2	1	9	10	6
slv	72	3894267	65	7	37	32	3	0	4	30	38
spa	20	2073675	16	4	4	12	4	5	10	3	2
srp	29	1087455	25	4	20	9	0	0	1	12	16

ELTeC – The Hungarian Corpus

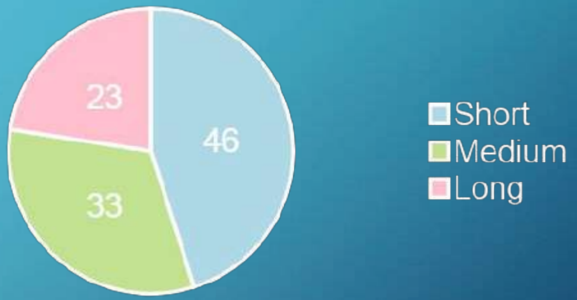
Title counts for each balance criterion



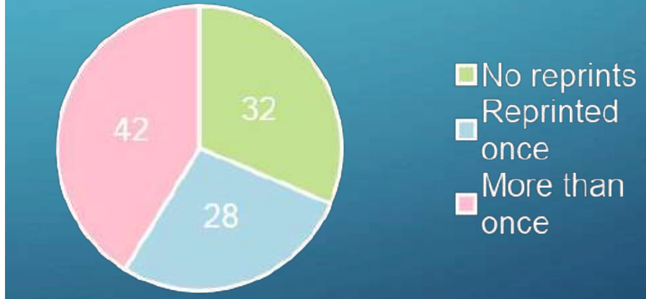
Date



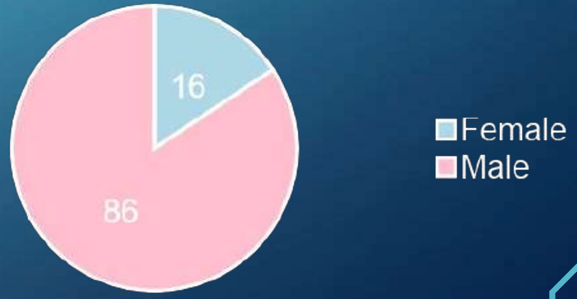
Length



Reprint Count

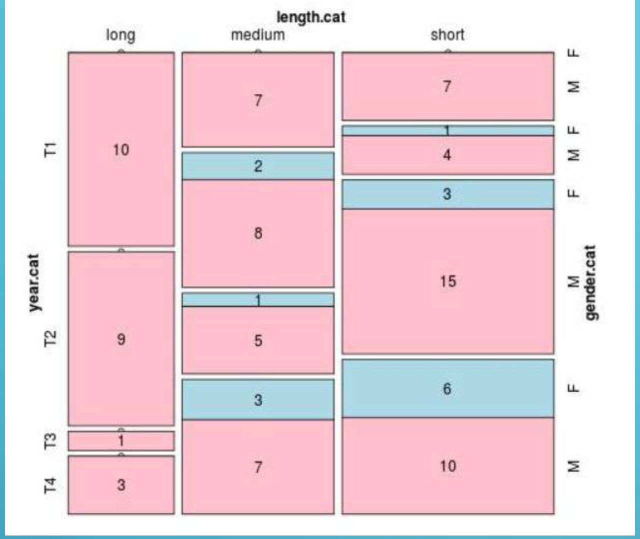


Author Gender

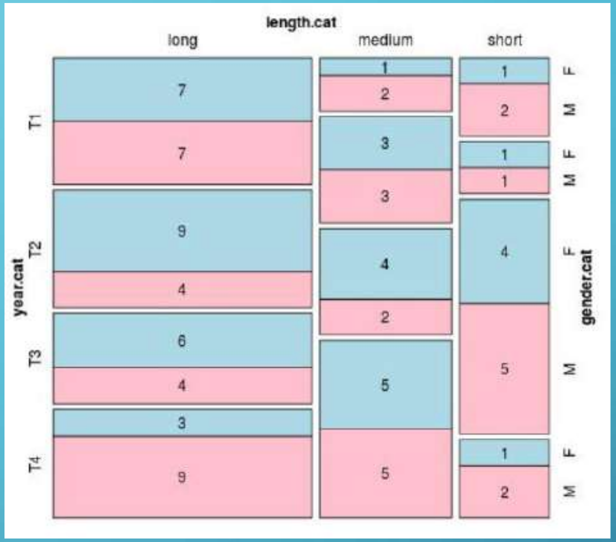


ELTeC – Comparisons

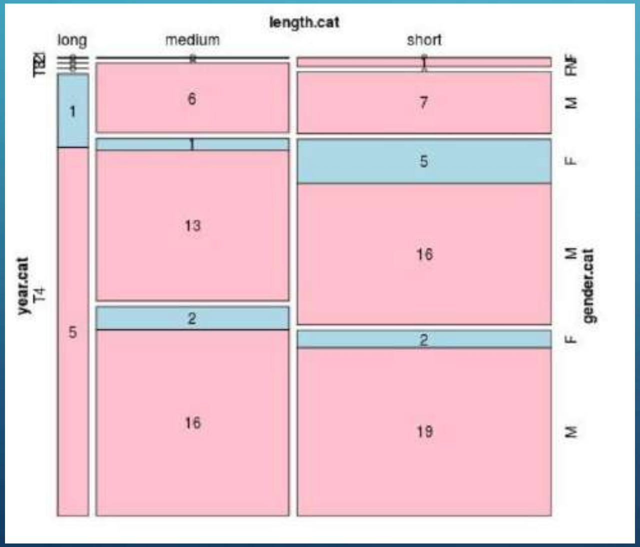
Hungarian



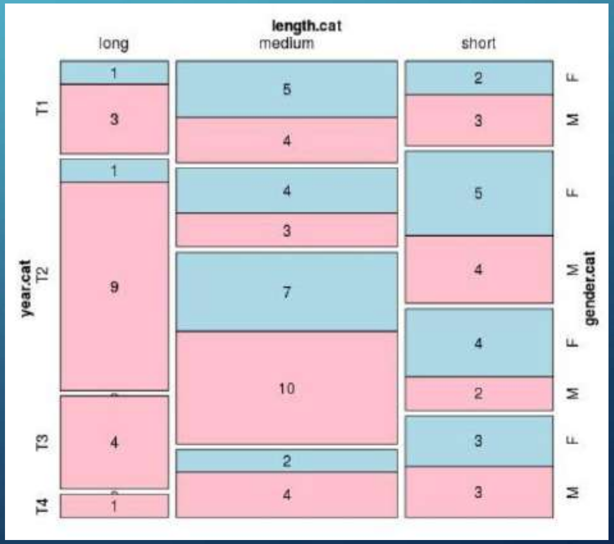
English



Slovenian

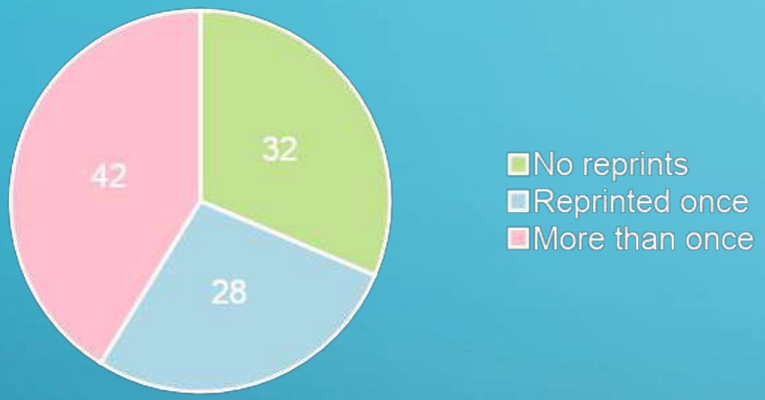


French

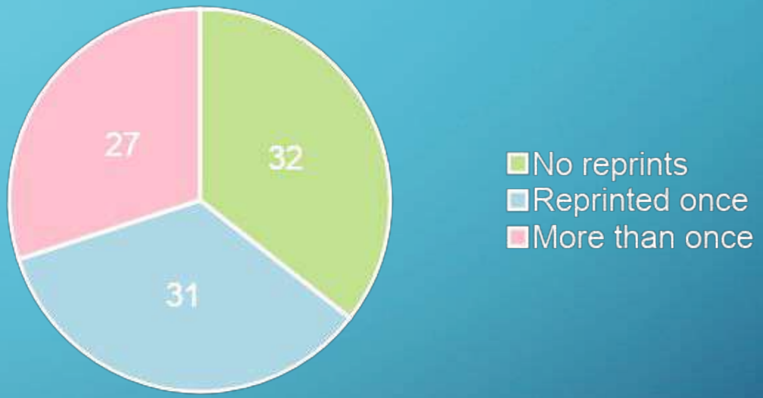


ELTeC – Comparisons

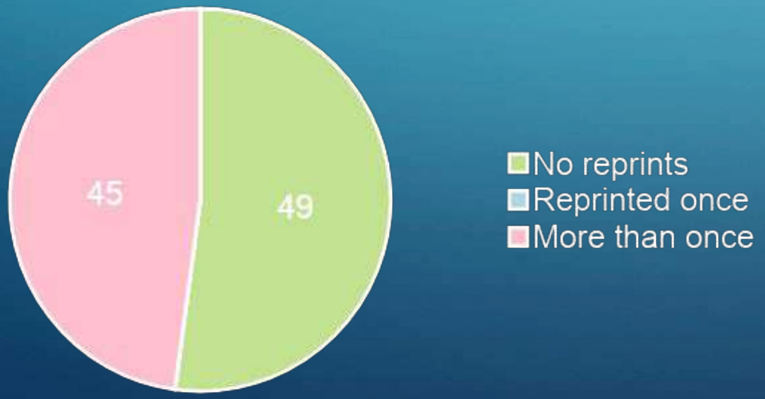
Reprint Count – Hungarian



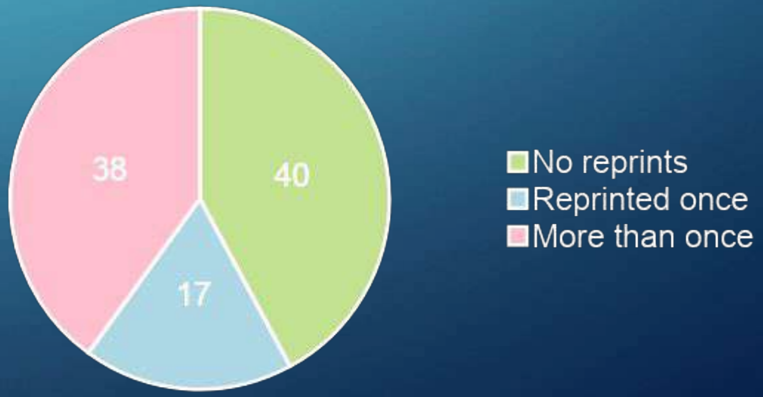
Reprint Count – English



Reprint Count – Slovenian

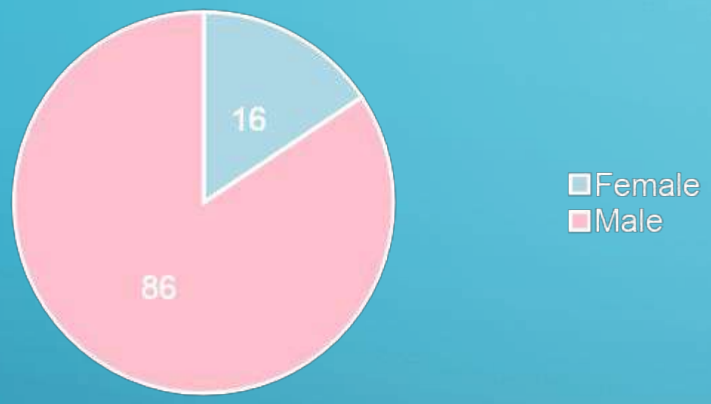


Reprint Count – French

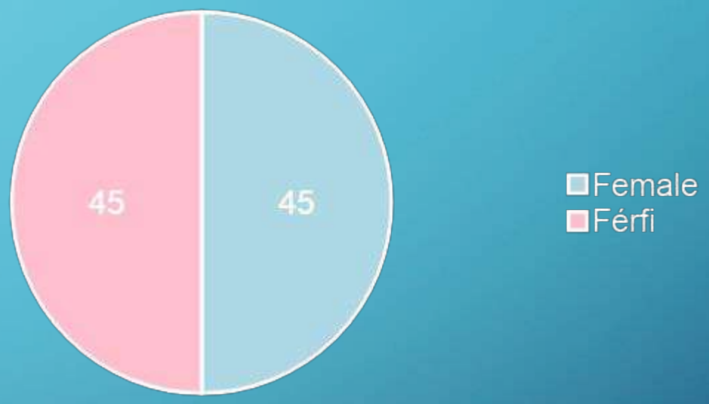


ELTeC – Comparisons

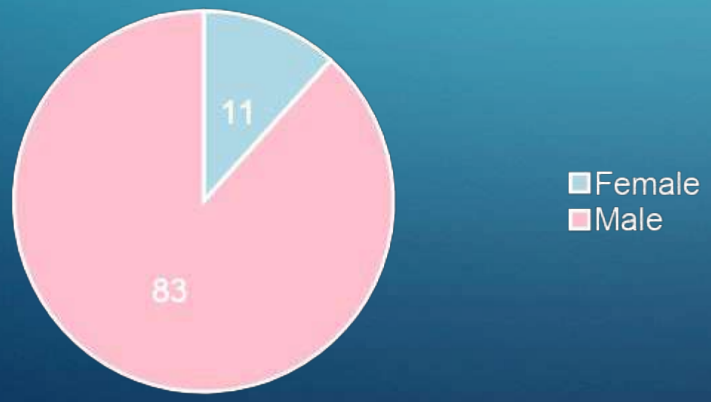
Author Gender – Hungarian



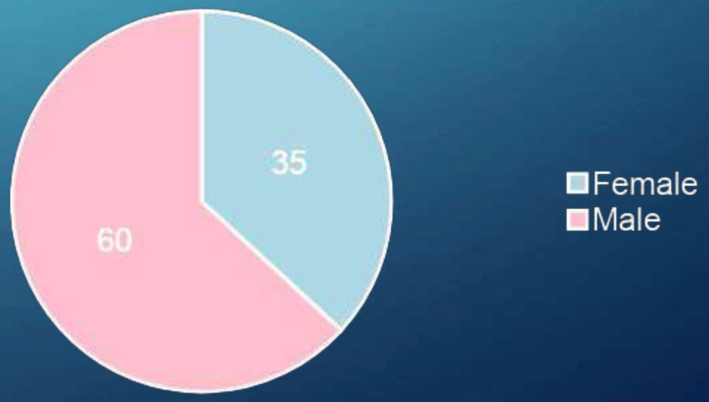
Author Gender – English



Author Gender - Slovenian



Author Gender - French

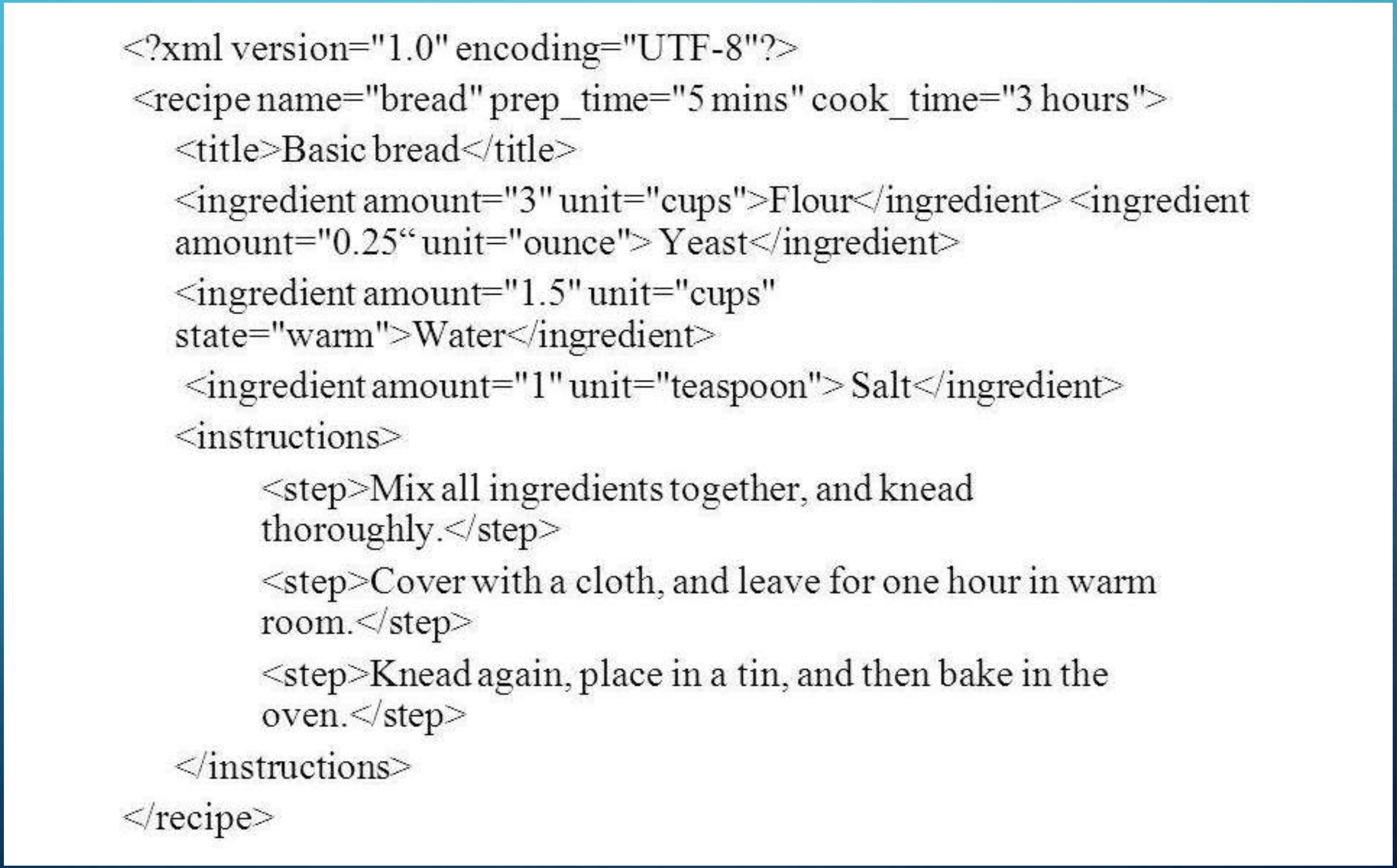


ELTeC – XML

- a markup language that defines a set of rules for encoding documents: **XML** (eXtensible Markup Language)
- Both human- and machine-readable
- Simplicity, generality, and usability across the Internet
- Metalanguage: describes an other language (=object language)
- Marking up different types of data
- Structured text and information sharing

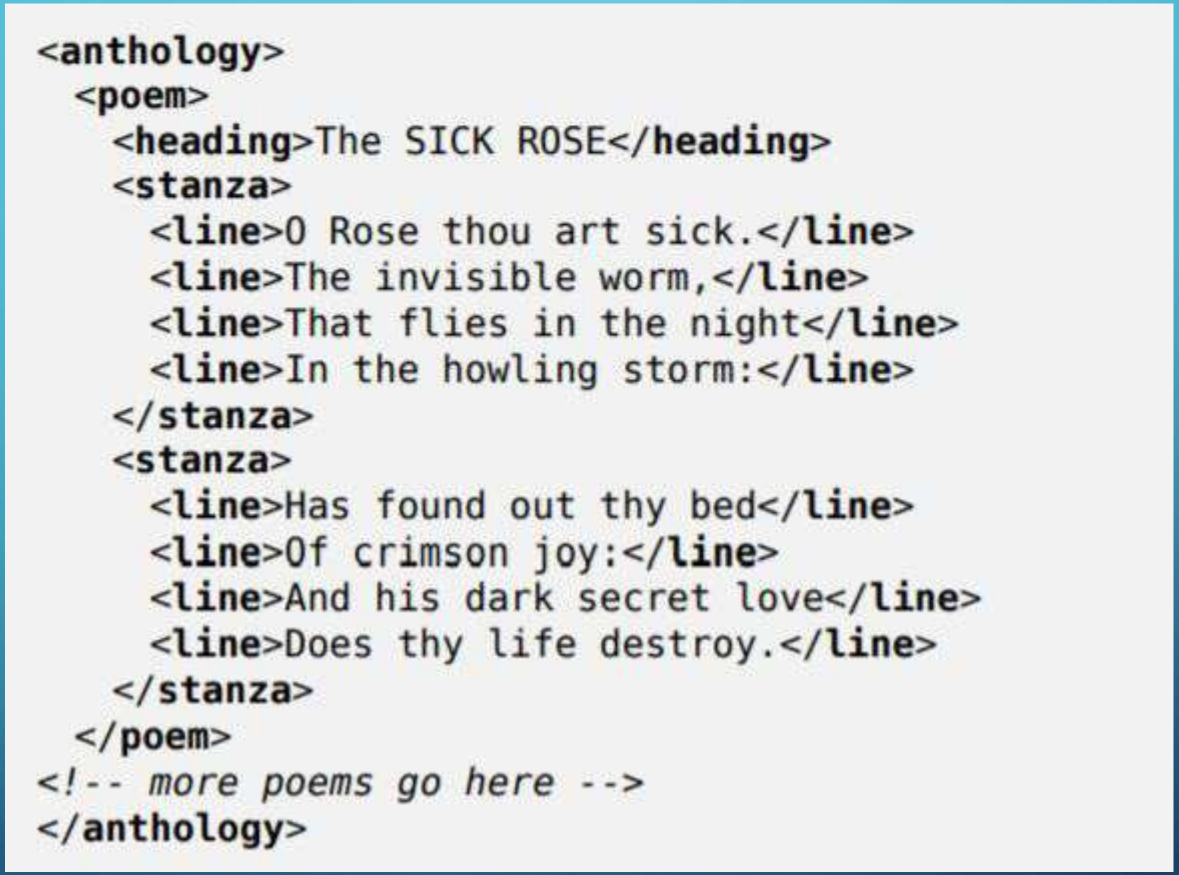
ELTeC – XML

ELTeC – XML



```
<?xml version="1.0" encoding="UTF-8"?>
 <recipe name="bread" prep_time="5 mins" cook_time="3 hours">
    <title>Basic bread</title>
    <ingredient amount="3" unit="cups">Flour</ingredient> <ingredient
    amount="0.25“ unit="ounce"> Yeast</ingredient>
    <ingredient amount="1.5" unit="cups"
    state="warm">Water</ingredient>
     <ingredient amount="1" unit="teaspoon"> Salt</ingredient>
    <instructions>
          <step>Mix all ingredients together, and knead
          thoroughly.</step>
          <step>Cover with a cloth, and leave for one hour in warm
          room.</step>
          <step>Knead again, place in a tin, and then bake in the
          oven.</step>
    </instructions>
</recipe>
```

ELTeC – XML

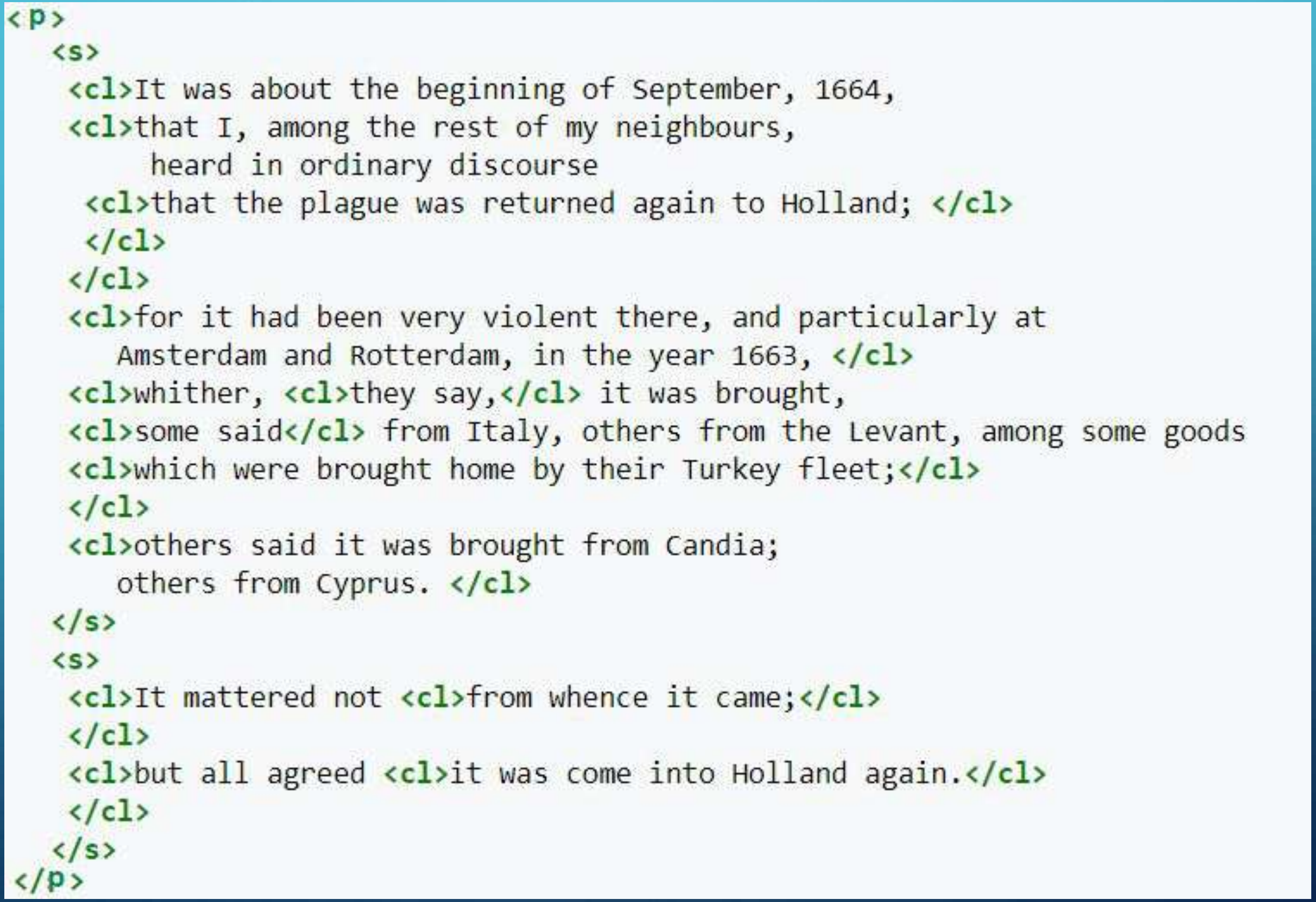


```
<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
<!-- more poems go here -->
</anthology>
```

ELTeC – TEI XML

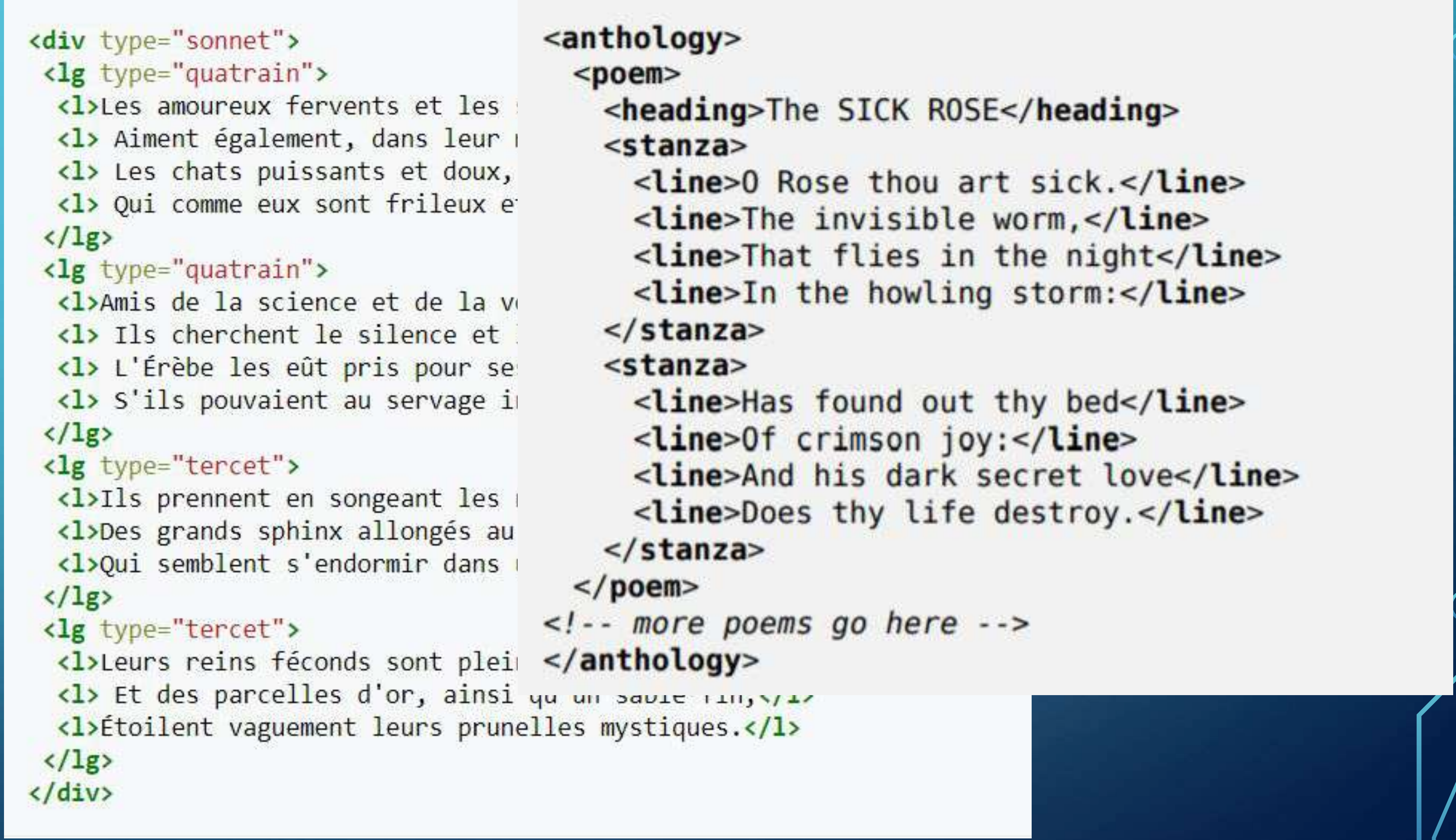
- Text Encoding Initiative
- A type of XML format used by Digital Humanities
- TEI Guidelines: standardize the XML markup language
- every tag and attribute are specified
- primarily semantic rather than presentational
- Over 500 different textual components and concepts (e.g. <persName>, <div>, <note>)

ELTeC – TEI XML



```
<p>
  <s>
   <cl>It was about the beginning of September, 1664,
   <cl>that I, among the rest of my neighbours,
        heard in ordinary discourse
    <cl>that the plague was returned again to Holland; </cl>
    </cl>
   </cl>
   <cl>for it had been very violent there, and particularly at
      Amsterdam and Rotterdam, in the year 1663, </cl>
   <cl>whither, <cl>they say,</cl> it was brought,
   <cl>some said</cl> from Italy, others from the Levant, among some goods
   <cl>which were brought home by their Turkey fleet;</cl>
   </cl>
   <cl>others said it was brought from Candia;
      others from Cyprus. </cl>
  </s>
  <s>
   <cl>It mattered not <cl>from whence it came;</cl>
   </cl>
   <cl>but all agreed <cl>it was come into Holland again.</cl>
   </cl>
  </s>
</p>
```

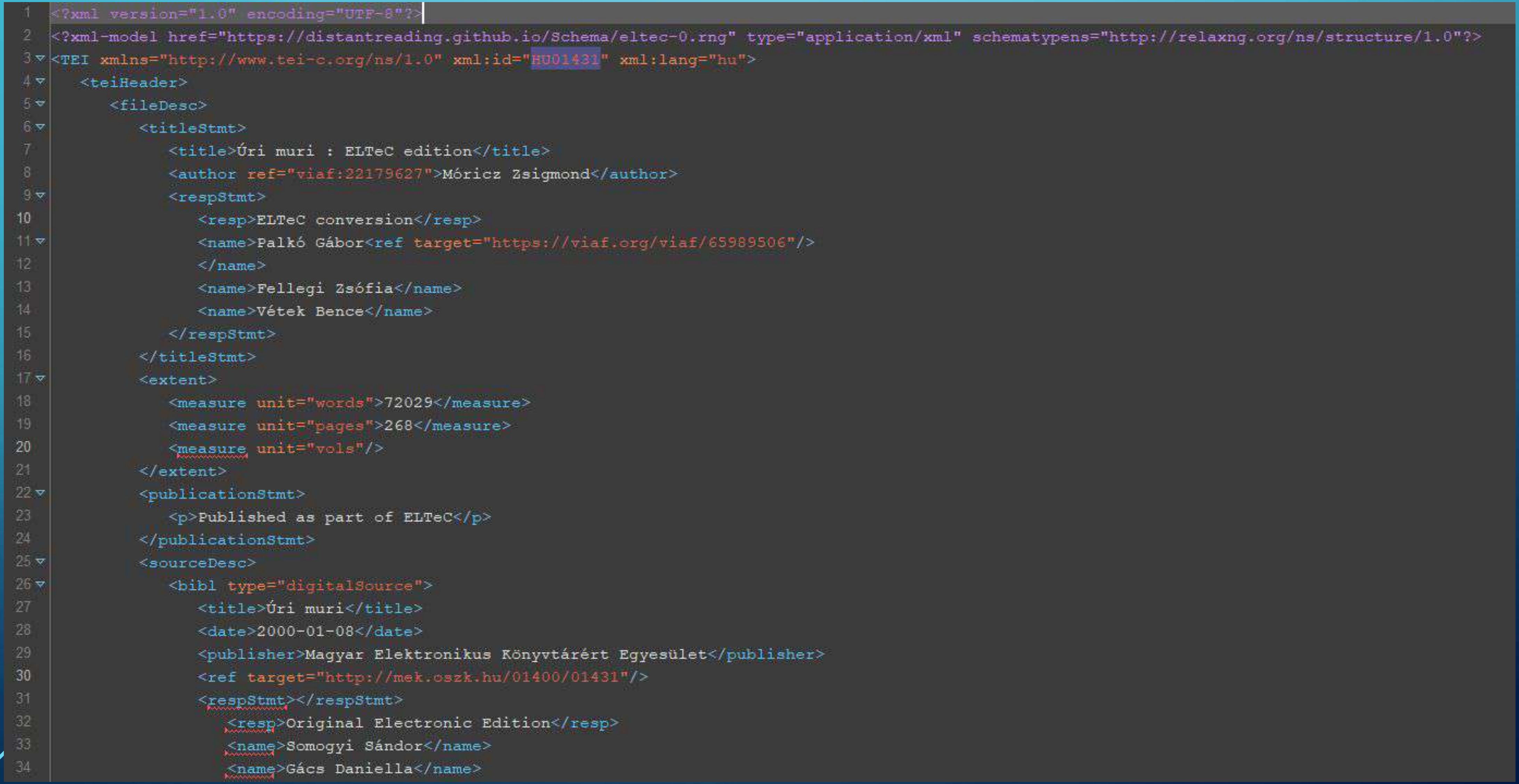
ELTeC – TEI XML



```
<div type="sonnet">
 <lg type="quatrain">
  <l>Les amoureux fervents et les
  <l> Aiment également, dans leur
  <l> Les chats puissants et doux,
  <l> Qui comme eux sont frileux e
 </lg>
 <lg type="quatrain">
  <l>Amis de la science et de la v
  <l> Ils cherchent le silence et
  <l> L'Érèbe les eût pris pour se
  <l> S'ils pouvaient au servage i
 </lg>
 <lg type="tercet">
  <l>Ils prennent en songeant les
  <l>Des grands sphinx allongés au
  <l>Qui semblent s'endormir dans
 </lg>
 <lg type="tercet">
  <l>Leurs reins féconds sont plei
  <l> Et des parcelles d'or, ainsi qu'un sable fin,</l>
  <l>Étoilent vaguement leurs prunelles mystiques.</l>
 </lg>
</div>
```

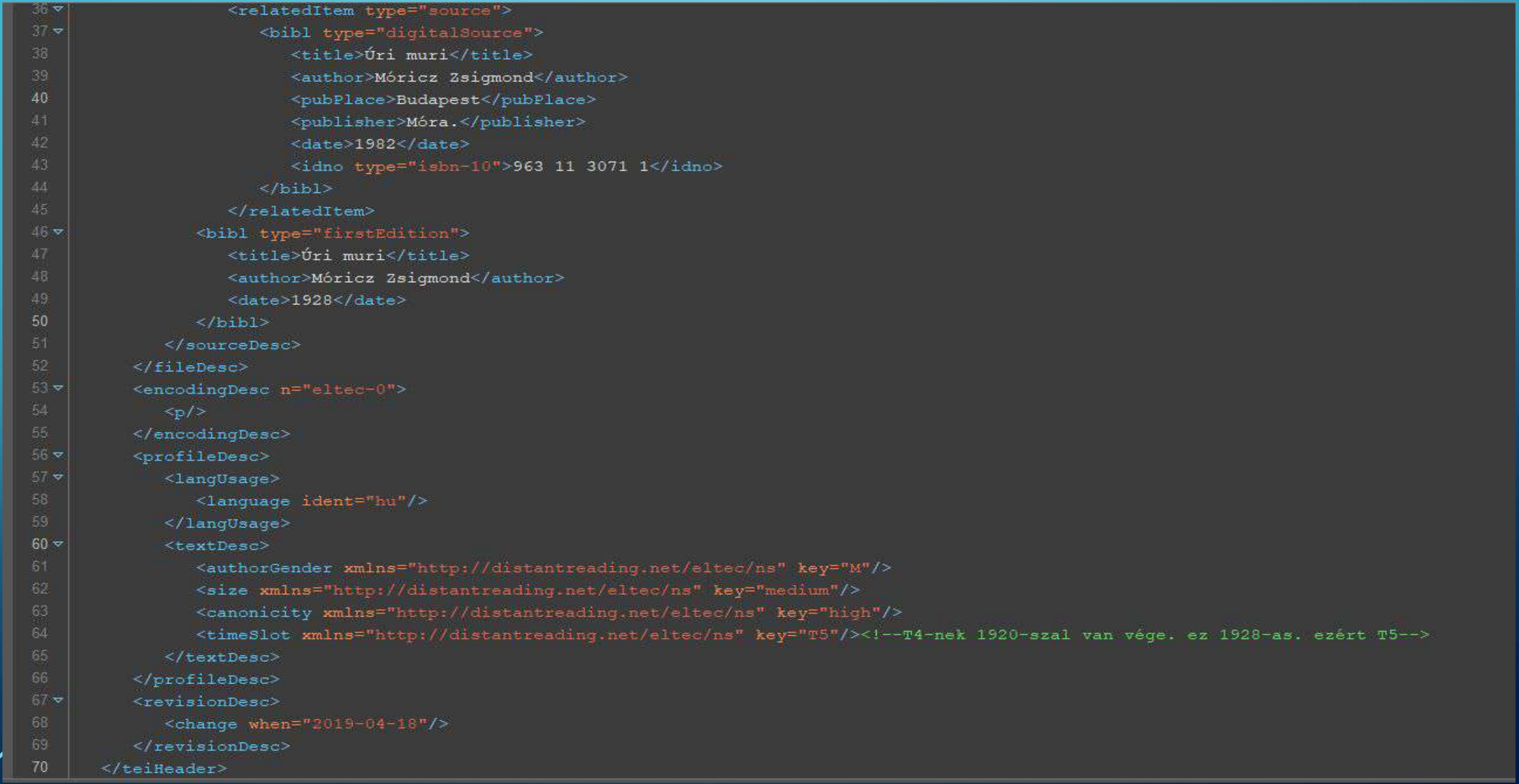
```
<anthology>
  <poem>
    <heading>The SICK ROSE</heading>
    <stanza>
      <line>O Rose thou art sick.</line>
      <line>The invisible worm,</line>
      <line>That flies in the night</line>
      <line>In the howling storm:</line>
    </stanza>
    <stanza>
      <line>Has found out thy bed</line>
      <line>Of crimson joy:</line>
      <line>And his dark secret love</line>
      <line>Does thy life destroy.</line>
    </stanza>
  </poem>
<!-- more poems go here -->
</anthology>
```

ELTeC – TEI XML



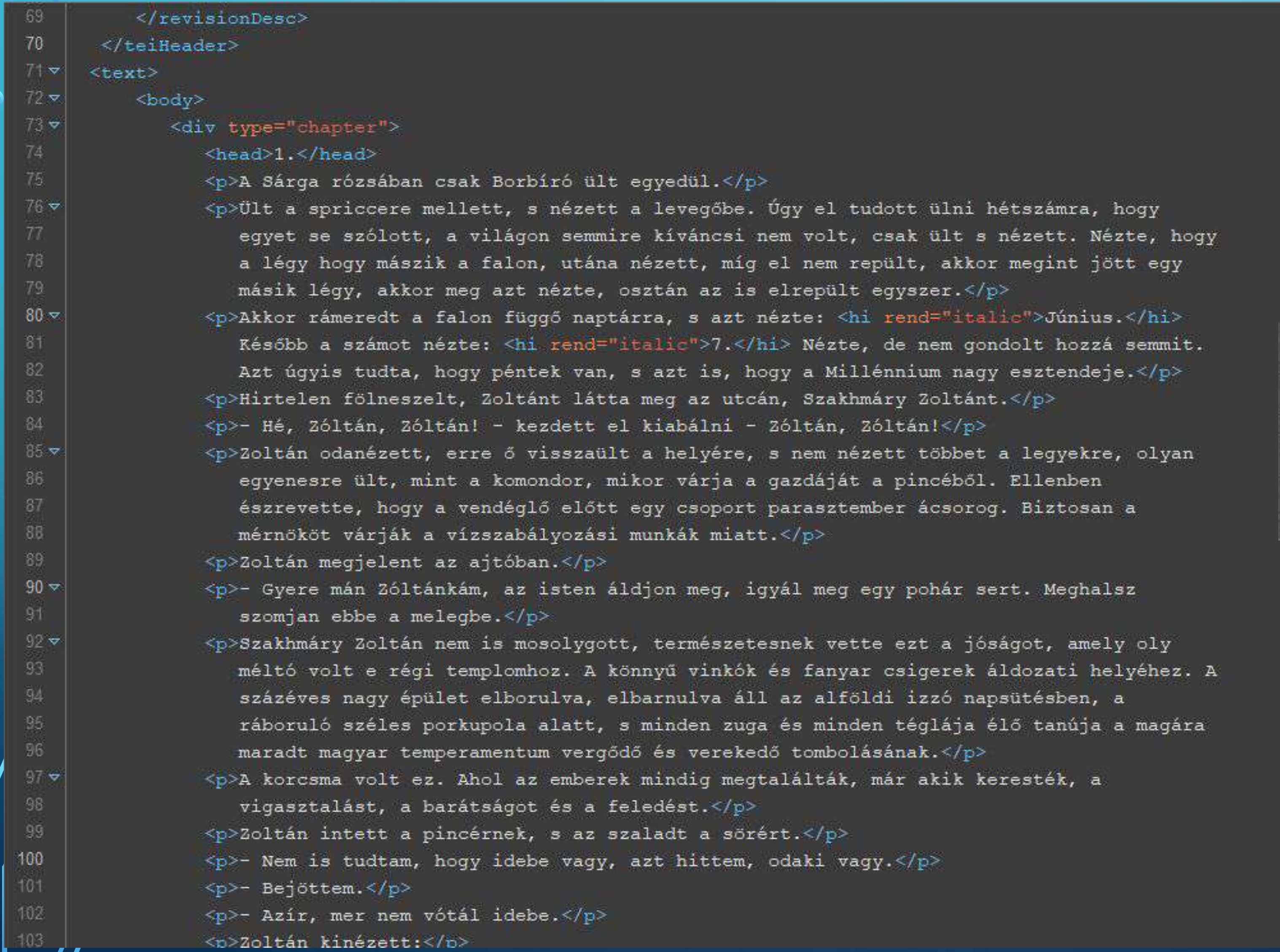
```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="https://distantreading.github.io/Schema/eltec-0.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="HU01431" xml:lang="hu">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title>Úri muri : ELTeC edition</title>
            <author ref="viaf:22179627">Móricz Zsigmond</author>
            <respStmt>
               <resp>ELTeC conversion</resp>
               <name>Palkó Gábor<ref target="https://viaf.org/viaf/65989506"/>
               </name>
               <name>Fellegi Zsófia</name>
               <name>Vétek Bence</name>
            </respStmt>
         </titleStmt>
         <extent>
            <measure unit="words">72029</measure>
            <measure unit="pages">268</measure>
            <measure unit="vols"/>
         </extent>
         <publicationStmt>
            <p>Published as part of ELTeC</p>
         </publicationStmt>
         <sourceDesc>
            <bibl type="digitalSource">
               <title>Úri muri</title>
               <date>2000-01-08</date>
               <publisher>Magyar Elektronikus Könyvtárért Egyesület</publisher>
               <ref target="http://mek.oszk.hu/01400/01431"/>
               <respStmt></respStmt>
                  <resp>Original Electronic Edition</resp>
                  <name>Somogyi Sándor</name>
                  <name>Gács Daniella</name>
```

ELTeC – TEI XML

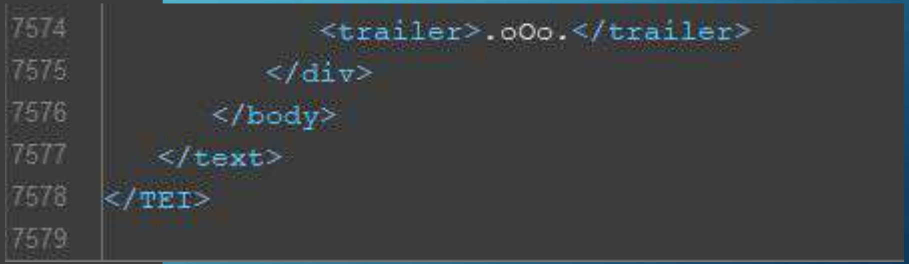


```
                  <relatedItem type="source">
                     <bibl type="digitalSource">
                        <title>Úri muri</title>
                        <author>Móricz Zsigmond</author>
                        <pubPlace>Budapest</pubPlace>
                        <publisher>Móra.</publisher>
                        <date>1982</date>
                        <idno type="isbn-10">963 11 3071 1</idno>
                     </bibl>
                  </relatedItem>
               <bibl type="firstEdition">
                  <title>Úri muri</title>
                  <author>Móricz Zsigmond</author>
                  <date>1928</date>
               </bibl>
            </sourceDesc>
         </fileDesc>
         <encodingDesc n="eltec-0">
            <p/>
         </encodingDesc>
         <profileDesc>
            <langUsage>
               <language ident="hu"/>
            </langUsage>
            <textDesc>
               <authorGender xmlns="http://distantreading.net/eltec/ns" key="M"/>
               <size xmlns="http://distantreading.net/eltec/ns" key="medium"/>
               <canonicity xmlns="http://distantreading.net/eltec/ns" key="high"/>
               <timeSlot xmlns="http://distantreading.net/eltec/ns" key="T5"/><!--T4-nek 1920-szal van vége. ez 1928-as. ezért T5-->
            </textDesc>
         </profileDesc>
         <revisionDesc>
            <change when="2019-04-18"/>
         </revisionDesc>
      </teiHeader>
```

ELTeC – TEI XML



```
      </revisionDesc>
   </teiHeader>
  <text>
      <body>
         <div type="chapter">
            <head>1.</head>
            <p>A Sárga rózsában csak Borbíró ült egyedül.</p>
            <p>Ült a spriccere mellett, s nézett a levegőbe. Úgy el tudott ülni hétszámra, hogy
               egyet se szólott, a világon semmire kíváncsi nem volt, csak ült s nézett. Nézte, hogy
               a légy hogy mászik a falon, utána nézett, míg el nem repült, akkor megint jött egy
               másik légy, akkor meg azt nézte, osztán az is elrepült egyszer.</p>
            <p>Akkor rámeredt a falon függő naptárra, s azt nézte: <hi rend="italic">Június.</hi>
               Később a számot nézte: <hi rend="italic">7.</hi> Nézte, de nem gondolt hozzá semmit.
               Azt úgyis tudta, hogy péntek van, s azt is, hogy a Millénnium nagy esztendeje.</p>
            <p>Hirtelen fölneszelt, Zoltánt látta meg az utcán, Szakhmáry Zoltánt.</p>
            <p>- Hé, Zóltán, Zóltán! - kezdett el kiabálni - Zóltán, Zóltán!</p>
            <p>Zoltán odanézett, erre ő visszaült a helyére, s nem nézett többet a legyekre, olyan
               egyenesre ült, mint a komondor, mikor várja a gazdáját a pincéből. Ellenben
               észrevette, hogy a vendéglő előtt egy csoport parasztember ácsorog. Biztosan a
               mérnököt várják a vízszabályozási munkák miatt.</p>
            <p>Zoltán megjelent az ajtóban.</p>
            <p>- Gyere mán Zóltánkám, az isten áldjon meg, igyál meg egy pohár sert. Meghalsz
               szomjan ebbe a melegbe.</p>
            <p>Szakhmáry Zoltán nem is mosolygott, természetesnek vette ezt a jóságot, amely oly
               méltó volt e régi templomhoz. A könnyű vinkók és fanyar csigerek áldozati helyéhez. A
               százéves nagy épület elborulva, elbarnulva áll az alföldi izzó napsütésben, a
               ráboruló széles porkupola alatt, s minden zuga és minden téglája élő tanúja a magára
               maradt magyar temperamentum vergődő és verekedő tombolásának.</p>
            <p>A korcsma volt ez. Ahol az emberek mindig megtalálták, már akik keresték, a
               vigasztalást, a barátságot és a feledést.</p>
            <p>Zoltán intett a pincérnek, s az szaladt a sörért.</p>
            <p>- Nem is tudtam, hogy idebe vagy, azt hittem, odaki vagy.</p>
            <p>- Bejöttem.</p>
            <p>- Azír, mer nem vótál idebe.</p>
            <p>Zoltán kinézett:</p>
```



```
            <trailer>.oOo.</trailer>
         </div>
      </body>
   </text>
</TEI>
```

Thanks for your attention!