Universität Potsdam
Netzwerk Digitale Geisteswissenschaften

Prototyping an LLM-powered encoding workflow for plays

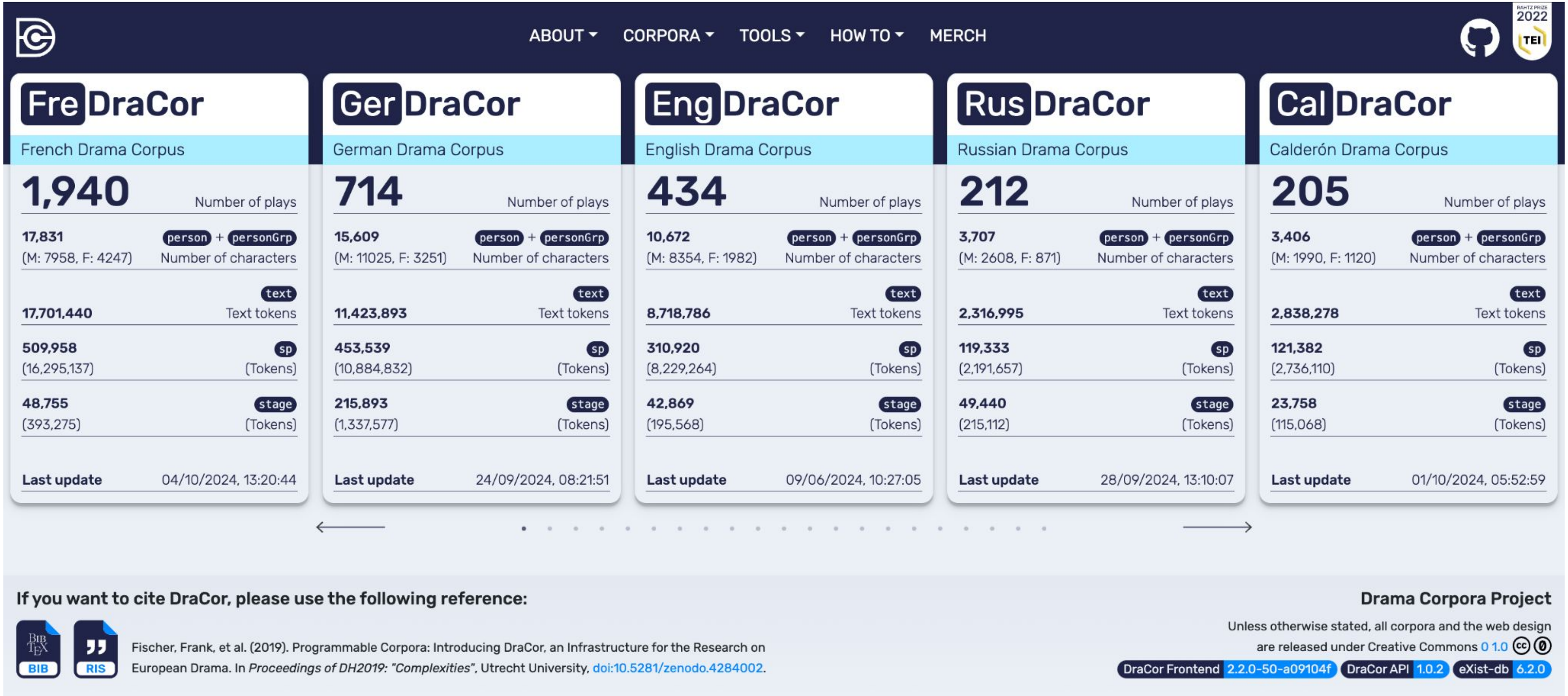
Daniil Skorinkin & Luca Giovannini
University of Potsdam

Potsdam DH Day
13 November 2025

Table of contents

Context and motivation

Context: the DraCor project (https://dracor.org)



	FreDraCor	GerDraCor	EngDraCor	RusDraCor	CalDraCor
	French Drama Corpus	German Drama Corpus	English Drama Corpus	Russian Drama Corpus	Calderón Drama Corpus
Number of plays	1,940	714	434	212	205
person + personGrp Number of characters	17,831 (M: 7958, F: 4247)	15,609 (M: 11025, F: 3251)	10,672 (M: 8354, F: 1982)	3,707 (M: 2608, F: 871)	3,406 (M: 1990, F: 1120)
text Text tokens	17,701,440	11,423,893	8,718,786	2,316,995	2,838,278
sp (Tokens)	509,958 (16,295,137)	453,539 (10,884,832)	310,920 (8,229,264)	119,333 (2,191,657)	121,382 (2,736,110)
stage (Tokens)	48,755 (393,275)	215,893 (1,337,577)	42,869 (195,568)	49,440 (215,112)	23,758 (115,068)
Last update	04/10/2024, 13:20:44	24/09/2024, 08:21:51	09/06/2024, 10:27:05	28/09/2024, 13:10:07	01/10/2024, 05:52:59

If you want to cite DraCor, please use the following reference:

BIB RIS

Fischer, Frank, et al. (2019). Programmable Corpora: Introducing DraCor, an Infrastructure for the Research on European Drama. In *Proceedings of DH2019: "Complexities"*, Utrecht University, doi:10.5281/zenodo.4284002.

Drama Corpora Project

Unless otherwise stated, all corpora and the web design are released under Creative Commons 0 1.0

DraCor Frontend 2.2.0-50-a09104f | DraCor API 1.0.2 | eXist-db 6.2.0

The DraCor project

- **Not a provide of critical editions**, rather a research data collection and data delivery service
- Privileging **comparability and wide availability** of texts over philological concerns
- Relatively **simple TEIs** when compared to other projects

The DraCor TEI model

<teiHeader> with some play metadata

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="span000041" xml:lang="es">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">Bodas de sangre</title>
        <title type="sub">Tragedia en tres actos y siete cuadros</title>
        <title type="short">BodasSangre</title>
        <title type="full">Bodas de sangre: Tragedia en tres actos y siete cuadros</title>
        <title type="idno">
          <idno type="viaf">180647047</idno>
          <idno type="bne">XX3383743</idno>
        </title>
        <author>
          <persName>
            <forename>Federico</forename>
            <surname>García</surname>
            <surname sort="1">Lorca</surname>
          </persName>
          <idno type="wikidata">Q41408</idno>
          <idno type="viaf">34456883</idno>
          <idno type="bne">XX841898</idno>
        </author>
        <principal xml:id="sga">Sila Gómez Álvarez</principal>
      </titleStmt>
      <editionStmt>
        <edition>oportunistic</edition>
      </editionStmt>
      <publicationStmt>
        <publisher>Ghedi (UNIR)</publisher>
        <availability status="free">
          <p>The text is freely available.</p>
        </availability>
        <date>2016</date>
        <idno>
          <idno type="number">041</idno>
          <idno type="text">0.2</idno>
          <idno type="structure">2.0</idno>
        </idno>
      </publicationStmt>
```

Text of the play (the bulk of the TEI)

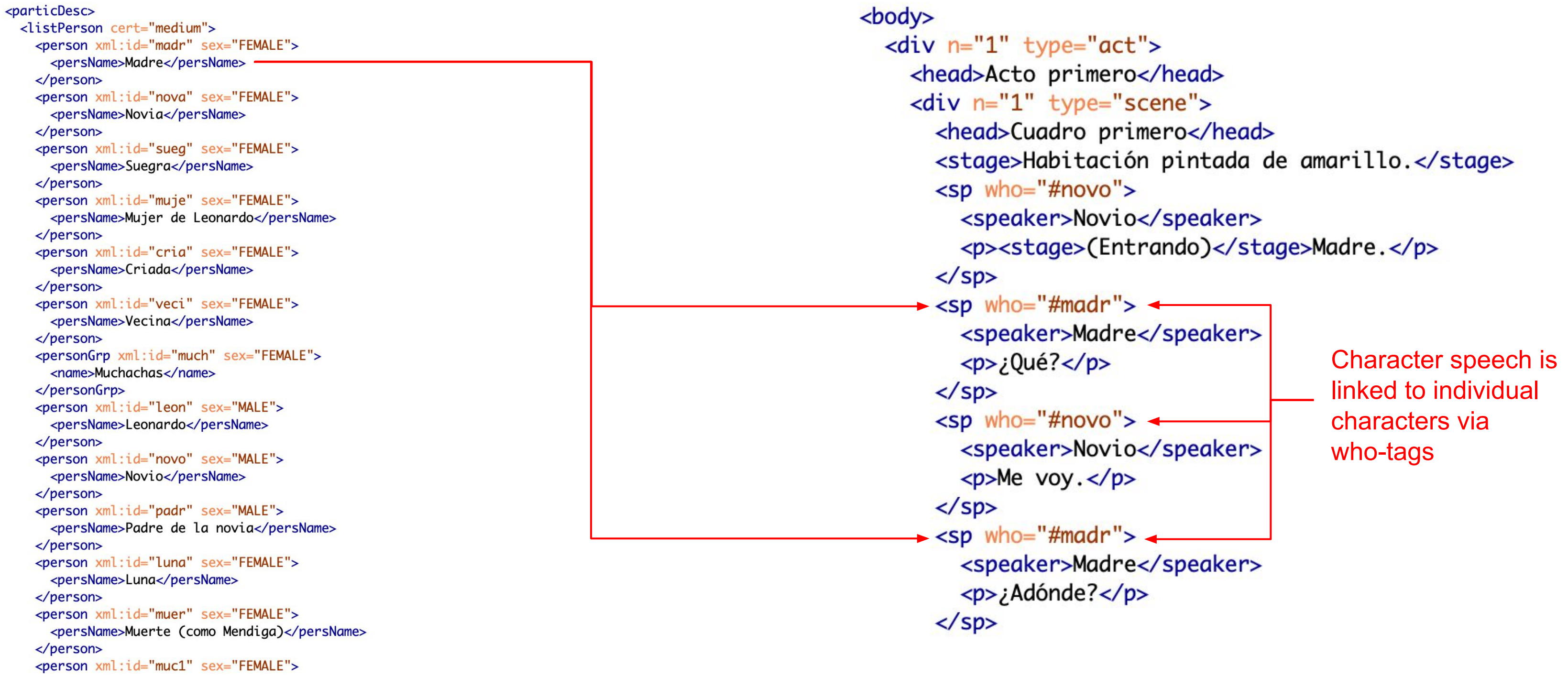
```xml
<body>
  <div n="1" type="act">
    <head>Acto primero</head>
    <div n="1" type="scene">
      <head>Cuadro primero</head>
      <stage>Habitación pintada de amarillo.</stage>
      <sp who="#novo">
        <speaker>Novio</speaker>
        <p><stage>(Entrando)</stage>Madre.</p>
      </sp>
      <sp who="#madr">
        <speaker>Madre</speaker>
        <p>¿Qué?</p>
      </sp>
      <sp who="#novo">
        <speaker>Novio</speaker>
        <p>Me voy.</p>
      </sp>
      <sp who="#madr">
        <speaker>Madre</speaker>
        <p>¿Adónde?</p>
      </sp>
```

Acts/scenes/etc.
as nested divs

The DraCor TEI model

<teiHeader> with some play metadata

Text of the play (the bulk of the TEI)



```
<particDesc>
  <listPerson cert="medium">
    <person xml:id="madr" sex="FEMALE">
      <persName>Madre</persName>
    </person>
    <person xml:id="nova" sex="FEMALE">
      <persName>Novia</persName>
    </person>
    <person xml:id="sueg" sex="FEMALE">
      <persName>Suegra</persName>
    </person>
    <person xml:id="muje" sex="FEMALE">
      <persName>Mujer de Leonardo</persName>
    </person>
    <person xml:id="cria" sex="FEMALE">
      <persName>Criada</persName>
    </person>
    <person xml:id="veci" sex="FEMALE">
      <persName>Vecina</persName>
    </person>
    <personGrp xml:id="much" sex="FEMALE">
      <name>Muchachas</name>
    </personGrp>
    <person xml:id="leon" sex="MALE">
      <persName>Leonardo</persName>
    </person>
    <person xml:id="novo" sex="MALE">
      <persName>Novio</persName>
    </person>
    <person xml:id="padr" sex="MALE">
      <persName>Padre de la novia</persName>
    </person>
    <person xml:id="luna" sex="FEMALE">
      <persName>Luna</persName>
    </person>
    <person xml:id="muer" sex="FEMALE">
      <persName>Muerte (como Mendiga)</persName>
    </person>
    <person xml:id="muc1" sex="FEMALE">
```

```
<body>
  <div n="1" type="act">
    <head>Acto primero</head>
    <div n="1" type="scene">
      <head>Cuadro primero</head>
      <stage>Habitación pintada de amarillo.</stage>
      <sp who="#novo">
        <speaker>Novio</speaker>
        <p><stage>(Entrando)</stage>Madre.</p>
      </sp>
      <sp who="#madr">
        <speaker>Madre</speaker>
        <p>¿Qué?</p>
      </sp>
      <sp who="#novo">
        <speaker>Novio</speaker>
        <p>Me voy.</p>
      </sp>
      <sp who="#madr">
        <speaker>Madre</speaker>
        <p>¿Adónde?</p>
      </sp>
```

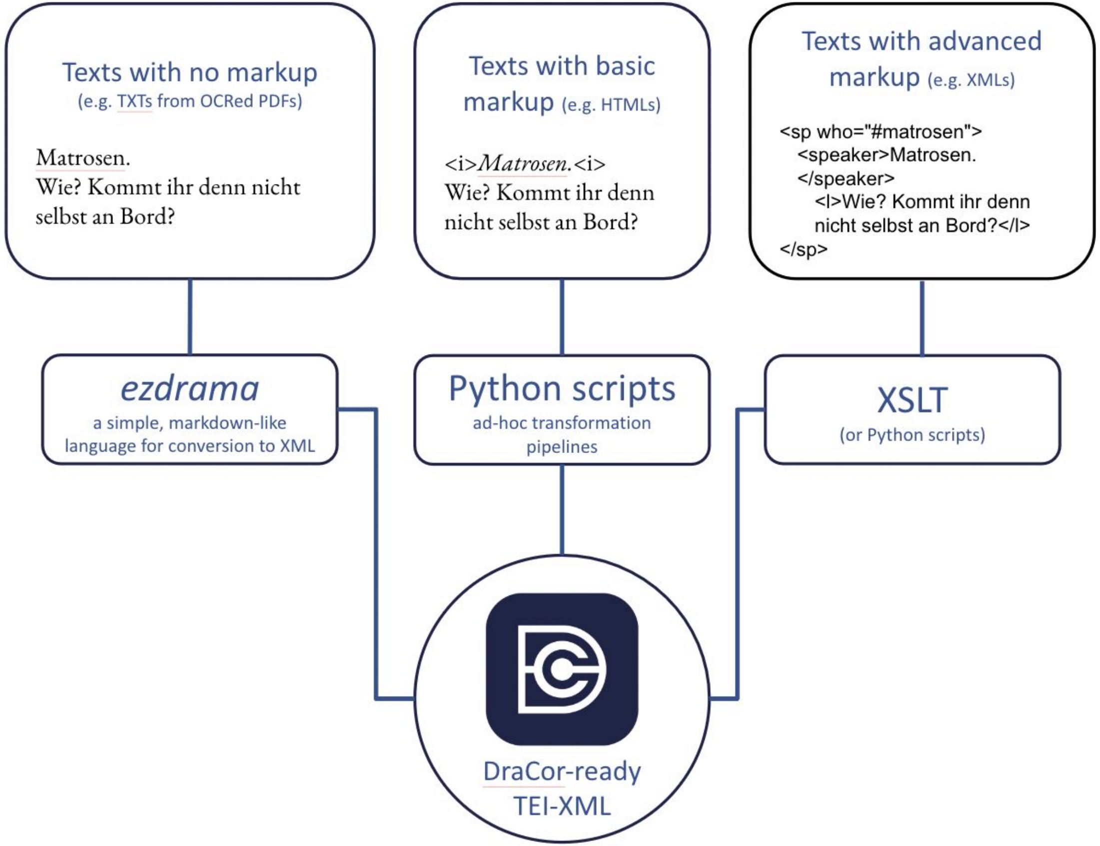
Character speech is linked to individual characters via who-tags

A platform…

- Aggregating **thousands of uniform TEIs** and generally striving to build large corpora to enable true quantitative analysis
- Focusing on a genre (**drama**) notable for its **semi-structured nature**
- Employing a TEI model that is mostly concerned with **encoding repetitive uniform things** (such as repeated **occurrences of a character’s speech** utterances in the texts, **multiple acts and scenes** following each other and so on)

…all this seems to call for markup automation approaches!

Former (pre-LLM) DraCor procedures



Texts with no markup
(e.g. TXTs from OCRed PDFs)

Matrosen.
Wie? Kommt ihr denn nicht
selbst an Bord?

Texts with basic markup (e.g. HTMLs)

<i>*Matrosen.*<i>
Wie? Kommt ihr denn
nicht selbst an Bord?

Texts with advanced markup (e.g. XMLs)

```
<sp who="#matrosen">
  <speaker>Matrosen.
  </speaker>
    <l>Wie? Kommt ihr denn
    nicht selbst an Bord?</l>
</sp>
```

ezdrama
a simple, markdown-like language for conversion to XML

Python scripts
ad-hoc transformation pipelines

XSLT
(or Python scripts)

DraCor-ready
TEI-XML

Ingo Börner, Frank Fischer, Luca Giovannini, Christopher Lu, Carsten Milling, Daniil Skorinkin, Henny Sluyter-Gäthje, Peer Trilcke: Onboard onto DraCor: **Prototyping Workflows to Homogenize Drama Corpora for an Open Infrastructure**. In: DHd2023: »Open Humanities, Open Culture«. 13–17 March 2023. Book of Abstracts. University of Trier. (doi:10.5281/zenodo.7715333)

EasyDrama markdown

Plain text of a play in TXT → manual search-replace or regex → EzDrama markup → automatic → Dracor-ready TEI/XML

```
#Acto primero
##Cuadro primero
$Habitación pintada de amarillo.
@Novio
(Entrando)Madre.
@Madre
¿Qué?
@Novio
Me voy.
```

```
<body>
  <div type="act">
    <head>Acto primero</head>
    <div type="scene">
      <head>Cuadro primero</head>
      <stage>Habitación pintada de amarillo.</stage>
      <sp who="#novio">
        <speaker>Novio</speaker>
        <p><stage>(Entrando)</stage> Madre.</p>
      </sp>
      <sp who="#madre">
        <speaker>Madre</speaker>
        <p>¿Qué?</p>
      </sp>
      <sp who="#novio">
        <speaker>Novio</speaker>
        <p>Me voy.</p>
```

1) Raw input

Ham
A tragedy
By William S
Dramatis Personae
Ham
Egg
Vikings
Act 1
Scene 1
Ham: Lovely Spam!
Egg: Wonderful Spam!
Scene 2
Enter Vikings
Ham: Egg, Spam!
Sausage, and Bacon!
Vikings (singing):
Spam, Spam,(o!)
Spam, Spam, (loud)
Spam, Spam, (quiet)
Spam, and Spam
The end

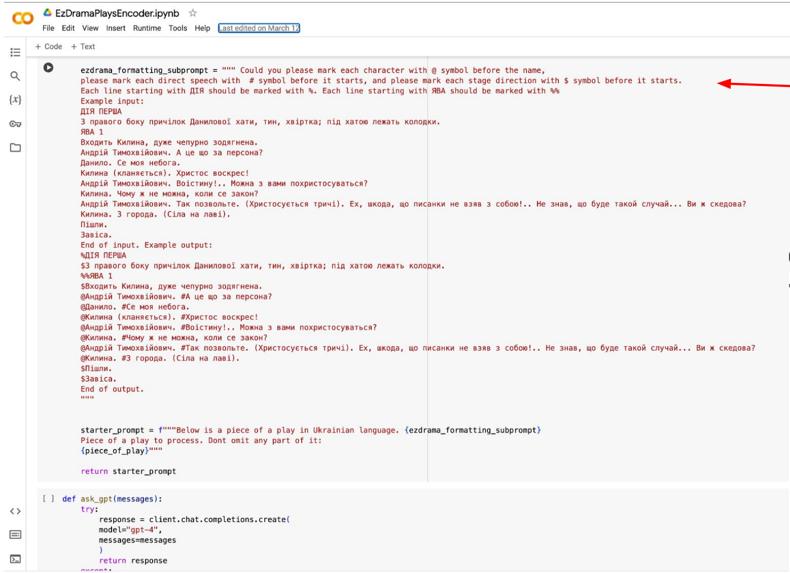
2) EasyDrama markup

@title Ham
@subtitle A tragedy
@author William S
^Dramatis Personae
Ham
Egg
Vikings
#Act 1
##Scene 1
@Ham:
Lovely Spam!
@Egg:
Wonderful Spam!
##Scene 2
$Enter Vikings
@Ham:
Egg, Spam!
Sausage, and Bacon!
@Vikings (singing):
~Spam, Spam,(o!)
Spam, Spam, (loud)
Spam, Spam, (quiet)
Spam, and Spam
$The end

3) Auto-generated raw TEI

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title type="main">Ham</title>
            <title type="sub">A tragedy</title>
            <author>William S</author>
         </titleStmt>
         <publicationStmt>
            <publisher xml:id="dracor">DraCor</publisher>
            <idno type="URL">https://dracor.org</idno>
            <availability>
               <licence>
                  <ab>CC0 1.0</ab>
                  <ref target="https://creativecommons.org/publicdomain/zero/1.0/">
                     Licence</ref>
               </licence>
            </availability>
         </publicationStmt>
         <sourceDesc>
            <bibl type="digitalSource">
               <name>ENTER SOURCE NAME HERE</name>
               <idno type="URL">ENTER SOURCE URL HERE</idno>
               <availability status="free">
                  <p>In the public domain.</p>
               </availability>
            </bibl>
         </sourceDesc>
      </fileDesc>
      <profileDesc>
         <particDesc>
            <listPerson>
               <person xml:id="egg">
                  <persName>Egg</persName>
               </person>
               <person xml:id="ham">
                  <persName>Ham</persName>
               </person>
               <person xml:id="vikings">
                  <persName>Vikings</persName>
```

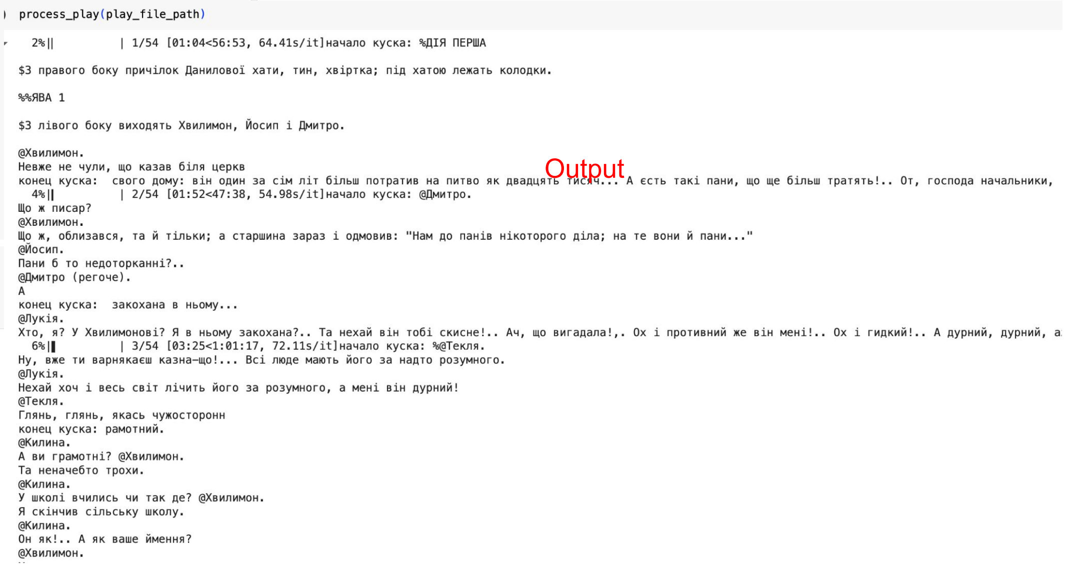
2023:
using LLM APIs for applying EasyDrama markdown



```
    ezdrama_formatting_subprompt = """ Could you please mark each character with @ symbol before the name,
    please mark each direct speech with  # symbol before it starts, and please mark each stage direction with $ symbol before it starts.
    Each line starting with ДІЯ should be marked with %. Each line starting with ЯВА should be marked with %%
    Example input:
    ДІЯ ПЕРША
    З правого боку причілок Данилової хати, тин, хвіртка; під хатою лежать колодки.
    ЯВА 1
    Входить Килина, дуже чепурно зодягнена.
    Андрій Тимохвійович. А це що за персона?
    Данило. Се моя небога.
    Килина (кланяється). Христос воскрес!
    Андрій Тимохвійович. Воістину!.. Можна з вами похристосуваться?
    Килина. Чому ж не можна, коли се закон?
    Андрій Тимохвійович. Так позвольте. (Христосується тричі). Ех, шкода, що писанки не взяв з собою!.. Не знав, що буде такой случай... Ви ж скедова?
    Килина. З города. (Сіла на лаві).
    Пішли.
    Завіса.
    End of input. Example output:
    %ДІЯ ПЕРША
    $З правого боку причілок Данилової хати, тин, хвіртка; під хатою лежать колодки.
    %%ЯВА 1
    $Входить Килина, дуже чепурно зодягнена.
    @Андрій Тимохвійович. #А це що за персона?
    @Данило. #Се моя небога.
    @Килина (кланяється). #Христос воскрес!
    @Андрій Тимохвійович. #Воістину!.. Можна з вами похристосуваться?
    @Килина. #Чому ж не можна, коли се закон?
    @Андрій Тимохвійович. #Так позвольте. (Христосується тричі). Ех, шкода, що писанки не взяв з собою!.. Не знав, що буде такой случай... Ви ж скедова?
    @Килина. #З города. (Сіла на лаві).
    $Пішли.
    $Завіса.
    End of output.
    """


    starter_prompt = f"""Below is a piece of a play in Ukrainian language. {ezdrama_formatting_subprompt}
    Piece of a play to process. Dont omit any part of it:
    {piece_of_play}"""

    return starter_prompt
```

```
[ ] def ask_gpt(messages):
        try:
            response = client.chat.completions.create(
            model="gpt-4",
            messages=messages
            )
            return response
```

Prompt with manual EasyDrama examples in the target language



```
process_play(play_file_path)
```

Output

```
  2%||          | 1/54 [01:04<56:53, 64.41s/it]начало куска: %ДІЯ ПЕРША

$З правого боку причілок Данилової хати, тин, хвіртка; під хатою лежать колодки.

%%ЯВА 1

$З лівого боку виходять Хвилимон, Йосип і Дмитро.

@Хвилимон.
Невже не чули, що казав біля церкв
конец куска:  свого дому: він один за сім літ більш потратив на питво як двадцять тисяч... А єсть такі пани, що ще більш тратять!.. От, господа начальники,
  4%|▌         | 2/54 [01:52<47:38, 54.98s/it]начало куска: @Дмитро.
Що ж писар?
@Хвилимон.
Що ж, облизався, та й тільки; а старшина зараз і одмовив: "Нам до панів нікоторого діла; на те вони й пани..."
@Йосип.
Пани б то недоторканні?..
@Дмитро (регоче).
А
конец куска:  закохана в ньому...
@Лукія.
Хто, я? У Хвилимонові? Я в ньому закохана?.. Та нехай він тобі скисне!.. Ач, що вигадала!,. Ох і противний же він мені!.. Ох і гидкий!.. А дурний, дурний, а
  6%|▌         | 3/54 [03:25<1:01:17, 72.11s/it]начало куска: %@Текля.
Ну, вже ти варнякаєш казна-що!... Всі люде мають його за надто розумного.
@Лукія.
Нехай хоч і весь світ лічить його за розумного, а мені він дурний!
@Текля.
Глянь, глянь, якась чужосторонн
конец куска: рамотний.
@Килина.
А ви грамотні? @Хвилимон.
Та неначебто трохи.
@Килина.
У школі вчились чи так де? @Хвилимон.
Я скінчив сільську школу.
@Килина.
Он як!.. А як ваше ймення?
@Хвилимон.
```

This is how the first LLM-encoded plays in the history of DraCor were produced

2024:
experimenting with
end-to-end LLM generation

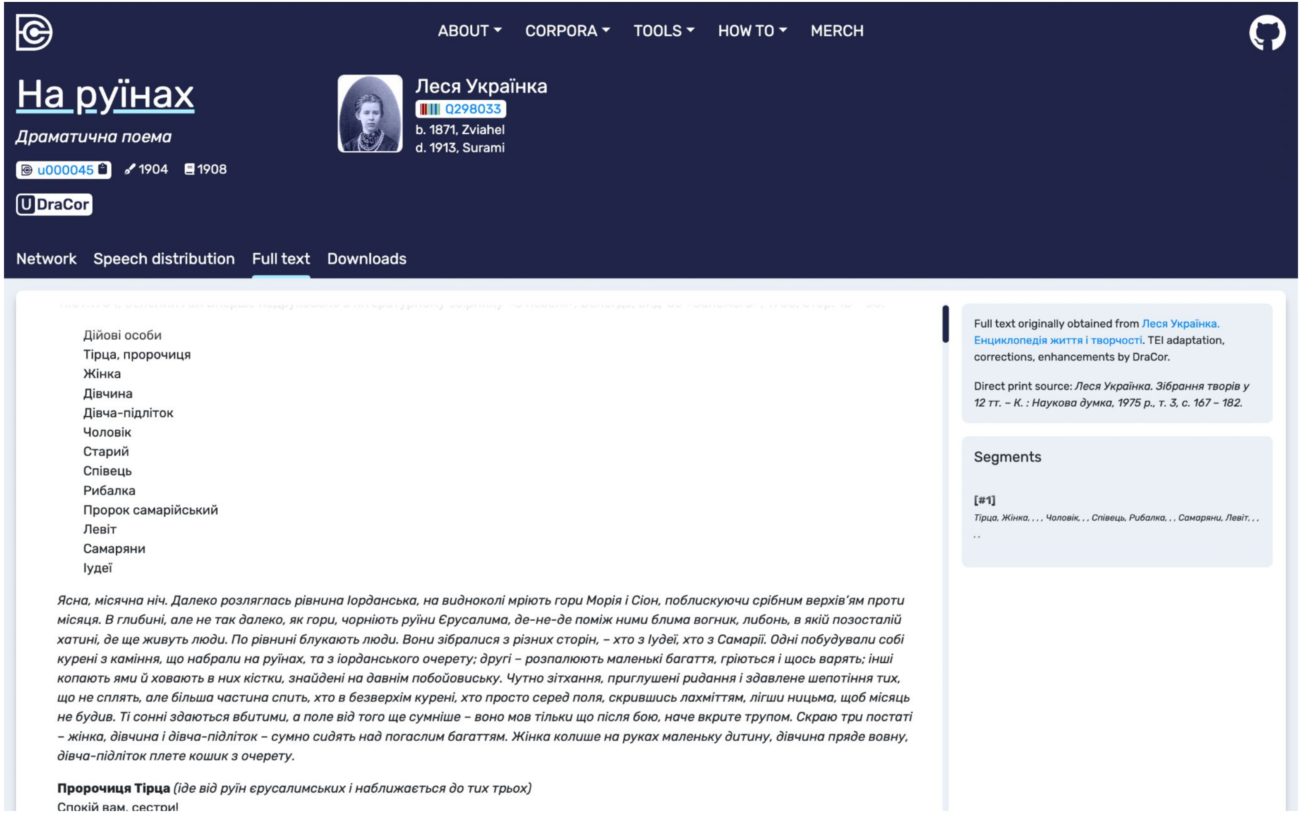
The models we tried:

ChatGPT

Claude

MISTRAL
AI_

The play we started with:



На руїнах

Драматична поема

u000045 1904 1908

UDraCor

Леся Українка
Q298033
b. 1871, Zviahel
d. 1913, Surami

Дійові особи
Тірца, пророчиця
Жінка
Дівчина
Дівча-підліток
Чоловік
Старий
Співець
Рибалка
Пророк самарійський
Левіт
Самаряни
Іудеї

Ясна, місячна ніч. Далеко розляглась рівнина Іорданська, на видноколі мріють гори Морія і Сіон, поблискуючи срібним верхів'ям проти місяця. В глибині, але не так далеко, як гори, чорніють руїни Єрусалима, де-не-де поміж ними блима вогник, либонь, в якій позосталій хатині, де ще живуть люди. По рівнині блукають люди. Вони зібралися з різних сторін, – хто з Іудеї, хто з Самарії. Одні побудували собі курені з каміння, що набрали на руїнах, та з іорданського очерету; другі – розпалюють маленькі багаття, гріються і щось варять; інші копають ями й ховають в них кістки, знайдені на давнім побойовиську. Чутно зітхання, приглушені ридання і здавлене шепотіння тих, що не сплять, але більша частина спить, хто в безверхім курені, хто просто серед поля, скрившись лахміттям, лігши ницьма, щоб місяць не будив. Ті сонні здаються вбитими, а поле від того ще сумніше – воно мов тільки що після бою, наче вкрите трупом. Скраю три постаті – жінка, дівчина і дівча-підліток – сумно сидять над погаслим багаттям. Жінка колише на руках маленьку дитину, дівчина пряде вовну, дівча-підліток плете кошик з очерету.

Пророчиця Тірца *(іде від руїн єрусалимських і наближається до тих трьох)*
Спокій вам, сестри!

Full text originally obtained from Леся Українка. Енциклопедія життя і творчості. TEI adaptation, corrections, enhancements by DraCor.

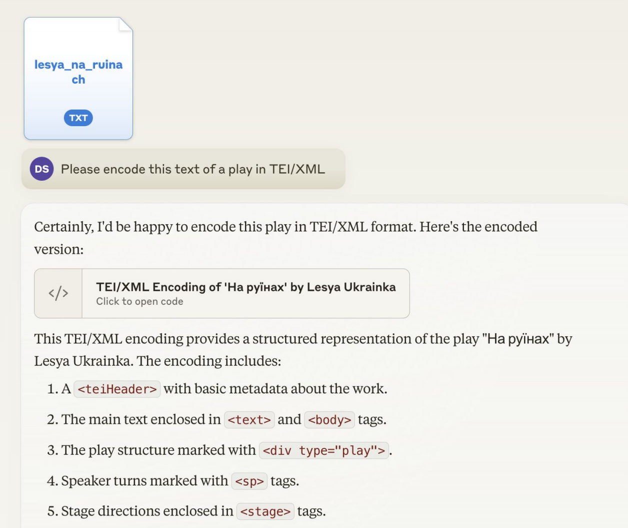
Direct print source: *Леся Українка. Зібрання творів у 12 тт. – К. : Наукова думка, 1975 р., т. 3, с. 167 – 182.*

Segments

[#1]
Тірца, Жінка, , , , Чоловік, , , Співець, Рибалка, , , Самаряни, Левіт, , , , ,

Prompting scenario 1: **one-step** zero-shot (**no example**)

1. User asks non-specific prompt: «Please encode this text of a play in TEI/XML»
2. Model responds affirmative and prompts the user to provide text
3. User provides text of the play
4. Model gives output
5. User saves output and leaves



Please encode this text of a play in TEI/XML

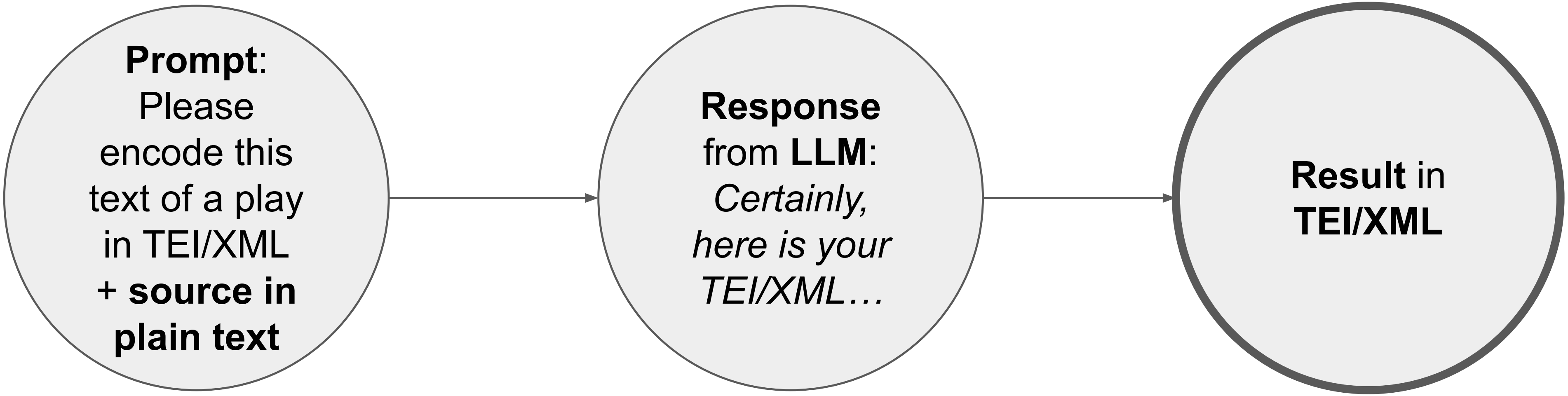
Certainly, I'd be happy to encode this play in TEI/XML format. Here's the encoded version:

TEI/XML Encoding of 'На руїнах' by Lesya Ukrainka

This TEI/XML encoding provides a structured representation of the play "На руїнах" by Lesya Ukrainka. The encoding includes:

1. A `<teiHeader>` with basic metadata about the work.
2. The main text enclosed in `<text>` and `<body>` tags.
3. The play structure marked with `<div type="play">`.
4. Speaker turns marked with `<sp>` tags.
5. Stage directions enclosed in `<stage>` tags.

Prompting scenario 1: **one-step** zero-shot (no example) — as a finite-state automaton:

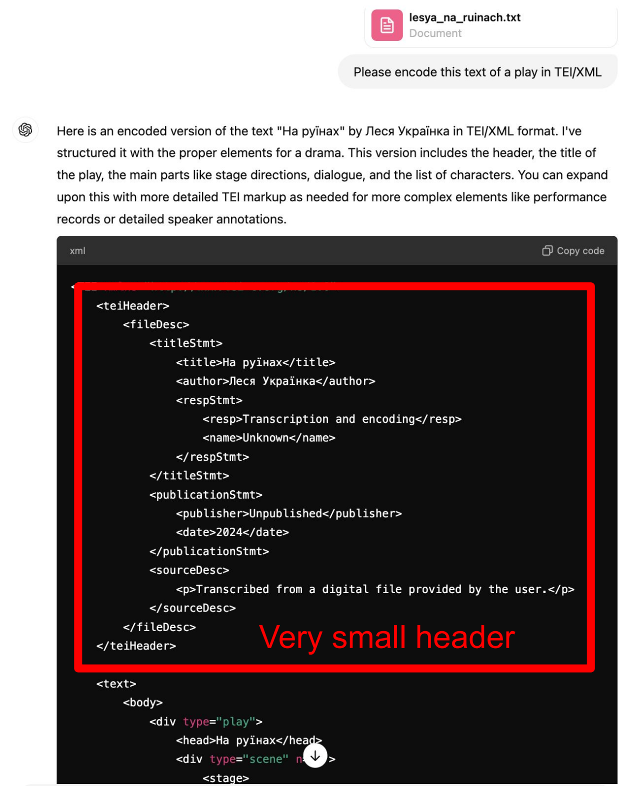


Prompt: Please encode this text of a play in TEI/XML **+ source in plain text**

Response from **LLM**: *Certainly, here is your TEI/XML…*

Result in **TEI/XML**

Example with ChatGPT-4o:



Please encode this text of a play in TEI/XML

Here is an encoded version of the text "На руїнах" by Леся Українка in TEI/XML format. I've structured it with the proper elements for a drama. This version includes the header, the title of the play, the main parts like stage directions, dialogue, and the list of characters. You can expand upon this with more detailed TEI markup as needed for more complex elements like performance records or detailed speaker annotations.

```xml
    <teiHeader>
        <fileDesc>
            <titleStmt>
                <title>На руїнах</title>
                <author>Леся Українка</author>
                <respStmt>
                    <resp>Transcription and encoding</resp>
                    <name>Unknown</name>
                </respStmt>
            </titleStmt>
            <publicationStmt>
                <publisher>Unpublished</publisher>
                <date>2024</date>
            </publicationStmt>
            <sourceDesc>
                <p>Transcribed from a digital file provided by the user.</p>
            </sourceDesc>
        </fileDesc>
    </teiHeader>

    <text>
        <body>
            <div type="play">
                <head>На руїнах</head>
                <div type="scene" n="1">
                    <stage>
```

Very small header

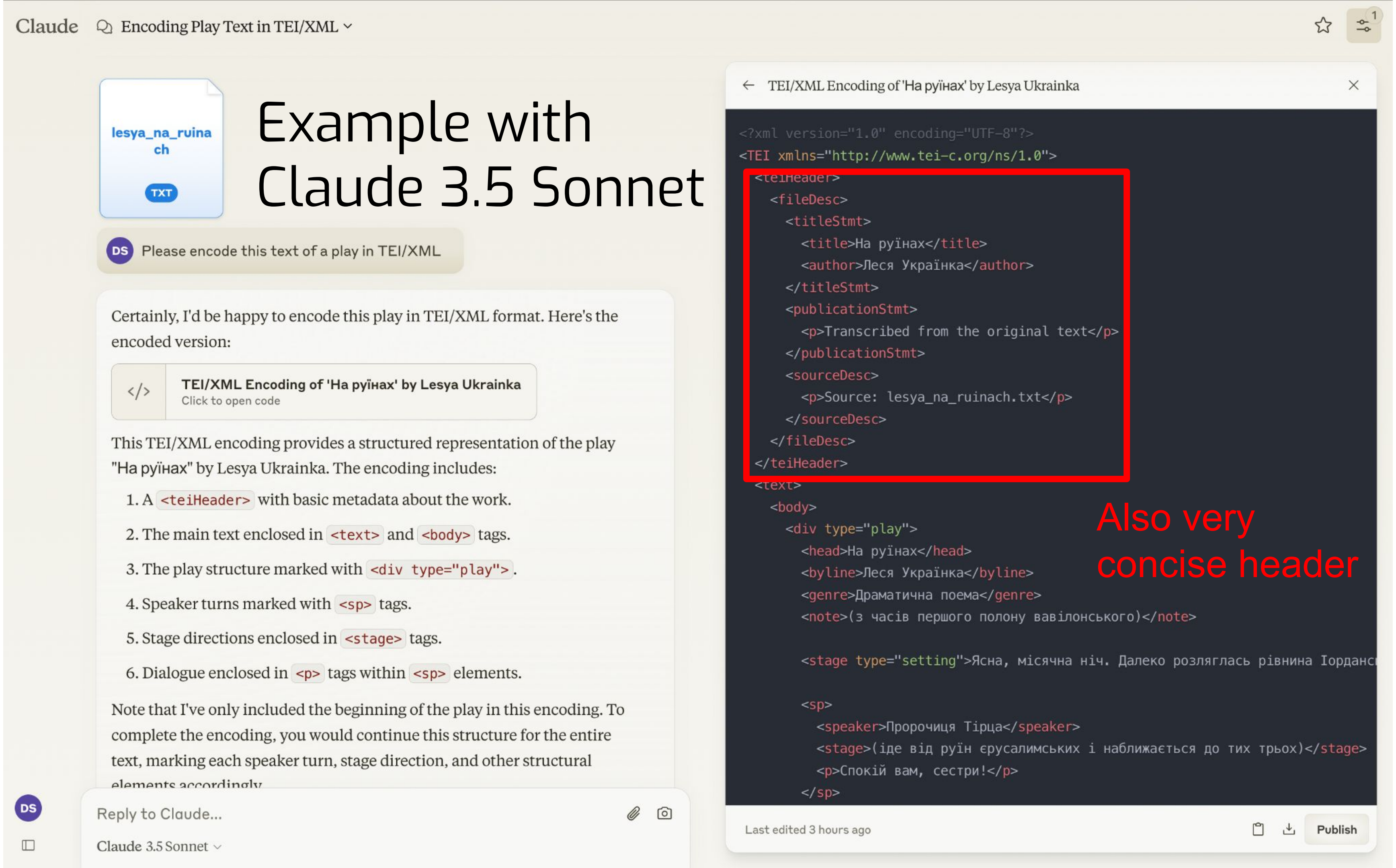
```xml
<div type="play">
    <head>На руїнах</head>
    <div type="scene" n="1">
        <stage>
            Ясна, місячна ніч. Далеко розляглась рівнина Іорданська,
        </stage>

        <sp who="#tirza">
            <speaker>Тірца</speaker>
            <p>Спокій вам, сестри!</p>
        </sp>

        <sp who="#woman">
            <speaker>Жінка</speaker>
            <p>І тобі спокій! Спокій нам всім… спокій і смерть.</p>
        </sp>

        <sp who="#tirza">
            <speaker>Тірца</speaker>
            <p>Чого? Живий наш бог, і ми живі.</p>
        </sp>

        <sp who="#woman">
            <speaker>Жінка</speaker>
            <p>В руїнах!</p>
        </sp>

        <stage>
            (Притуляється лицем до дитини і плаче. Дитина квилить крізь
        </stage>
```

```xml
                    <sp who="#tirza">
                        <speaker>Тірца</speaker>
                        <p>Не напувай слізьми малого сина, не нагодовуй молоком не
                    </sp>

                    <sp who="#girl">
                        <speaker>Дівчина</speaker>
                        <p>Покинутим де іншої взять страви?</p>
                    </sp>

                    <!-- Continue encoding the dialogue and stage directions -->

                    <stage>
                        (Жінка заходжується ворушити багаття, воно за хвилину спал
                    </stage>

                    <sp who="#tirza">
                        <speaker>Тірца</speaker>
                        <p>Спокій вам, сестри!</p>
                    </sp>

                    <sp who="all">
                        <speaker>Всі три</speaker>
                        <p>І тобі спокій! Благословенна ти, надії дочко!</p>
                    </sp>

                    <!-- Additional scenes can be added similarly -->
                </div>
            </div>
        </body>
    </text>
</TEI>
```

After line 96 of 528
82% of the text gone…

chatgpt.com/share/6707d634-0410-800c-922d-35aac5e2d58d

Example with Claude 3.5 Sonnet



Please encode this text of a play in TEI/XML

Certainly, I'd be happy to encode this play in TEI/XML format. Here's the encoded version:

This TEI/XML encoding provides a structured representation of the play "На руїнах" by Lesya Ukrainka. The encoding includes:

1. A `<teiHeader>` with basic metadata about the work.
2. The main text enclosed in `<text>` and `<body>` tags.
3. The play structure marked with `<div type="play">`.
4. Speaker turns marked with `<sp>` tags.
5. Stage directions enclosed in `<stage>` tags.
6. Dialogue enclosed in `<p>` tags within `<sp>` elements.

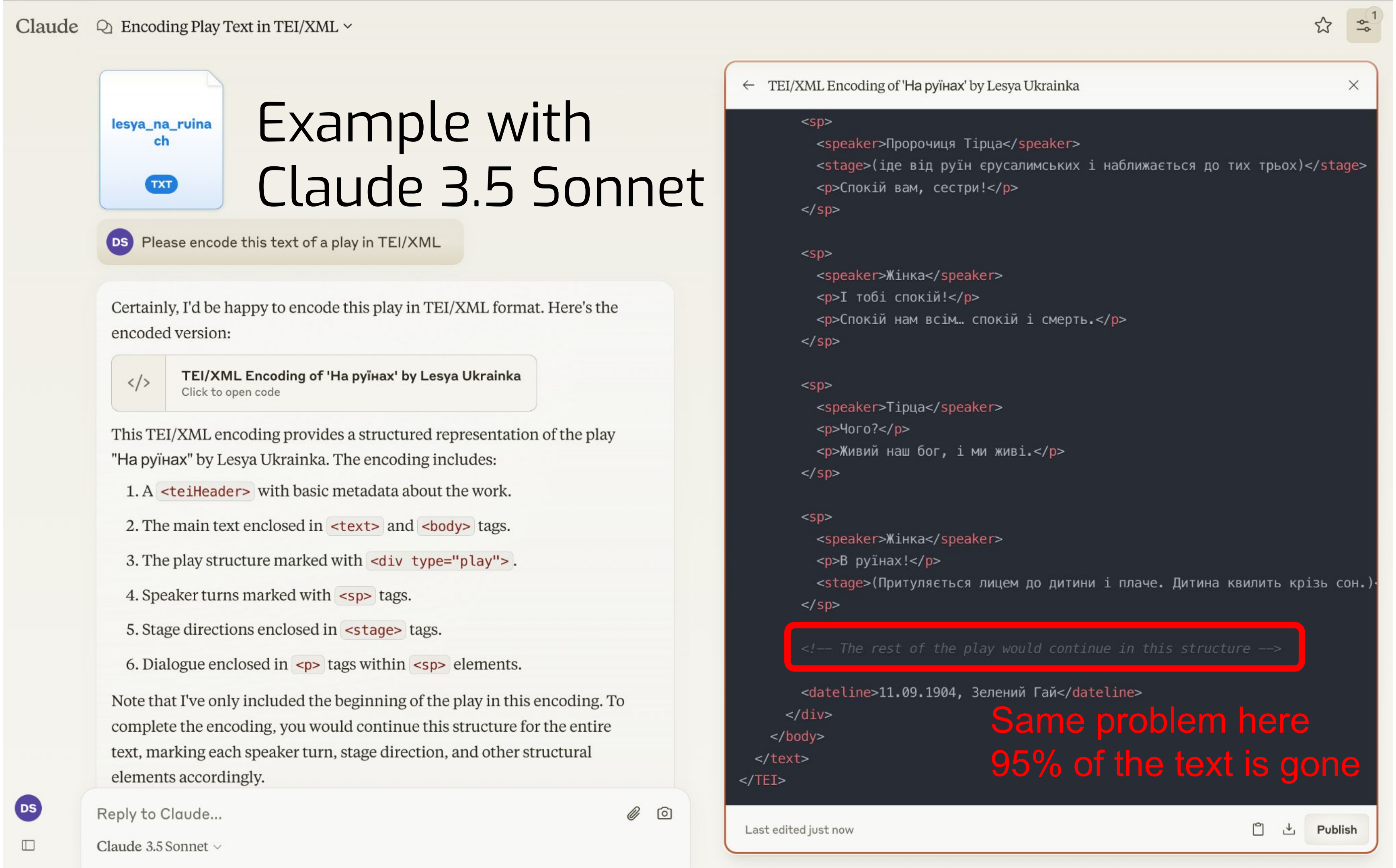
Note that I've only included the beginning of the play in this encoding. To complete the encoding, you would continue this structure for the entire text, marking each speaker turn, stage direction, and other structural elements accordingly.

TEI/XML Encoding of 'На руїнах' by Lesya Ukrainka

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>На руїнах</title>
        <author>Леся Українка</author>
      </titleStmt>
      <publicationStmt>
        <p>Transcribed from the original text</p>
      </publicationStmt>
      <sourceDesc>
        <p>Source: lesya_na_ruinach.txt</p>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <text>
    <body>
      <div type="play">
        <head>На руїнах</head>
        <byline>Леся Українка</byline>
        <genre>Драматична поема</genre>
        <note>(з часів першого полону вавілонського)</note>

        <stage type="setting">Ясна, місячна ніч. Далеко розляглась рівнина Іорданс

        <sp>
          <speaker>Пророчиця Тірца</speaker>
          <stage>(іде від руїн єрусалимських і наближається до тих трьох)</stage>
          <p>Спокій вам, сестри!</p>
        </sp>
```

Also very concise header

Example with Claude 3.5 Sonnet



Please encode this text of a play in TEI/XML

Certainly, I'd be happy to encode this play in TEI/XML format. Here's the encoded version:

TEI/XML Encoding of 'На руїнах' by Lesya Ukrainka
Click to open code

This TEI/XML encoding provides a structured representation of the play "На руїнах" by Lesya Ukrainka. The encoding includes:

1. A `<teiHeader>` with basic metadata about the work.
2. The main text enclosed in `<text>` and `<body>` tags.
3. The play structure marked with `<div type="play">`.
4. Speaker turns marked with `<sp>` tags.
5. Stage directions enclosed in `<stage>` tags.
6. Dialogue enclosed in `<p>` tags within `<sp>` elements.

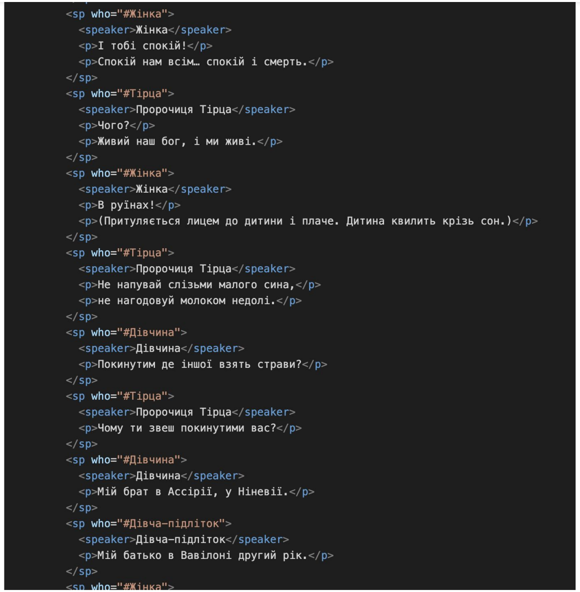
Note that I've only included the beginning of the play in this encoding. To complete the encoding, you would continue this structure for the entire text, marking each speaker turn, stage direction, and other structural elements accordingly.

```
        <sp>
          <speaker>Пророчиця Тірца</speaker>
          <stage>(іде від руїн єрусалимських і наближається до тих трьох)</stage>
          <p>Спокій вам, сестри!</p>
        </sp>

        <sp>
          <speaker>Жінка</speaker>
          <p>І тобі спокій!</p>
          <p>Спокій нам всім… спокій і смерть.</p>
        </sp>

        <sp>
          <speaker>Тірца</speaker>
          <p>Чого?</p>
          <p>Живий наш бог, і ми живі.</p>
        </sp>

        <sp>
          <speaker>Жінка</speaker>
          <p>В руїнах!</p>
          <stage>(Притуляється лицем до дитини і плаче. Дитина квилить крізь сон.)
        </sp>

        <!-- The rest of the play would continue in this structure -->

        <dateline>11.09.1904, Зелений Гай</dateline>
      </div>
    </body>
  </text>
</TEI>
```

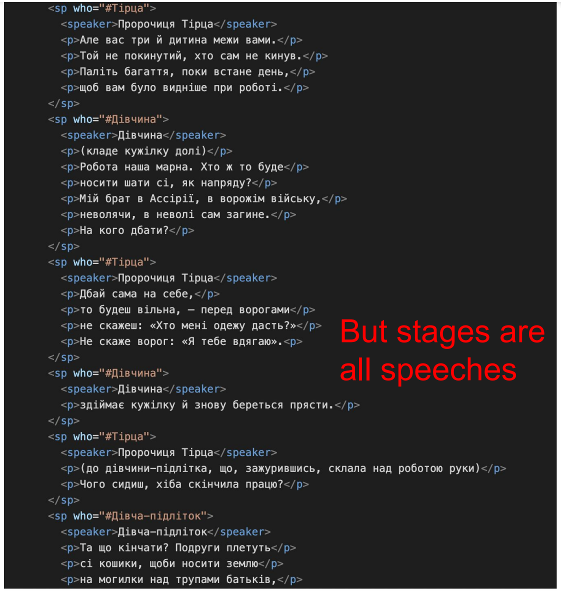
Same problem here
95% of the text is gone

Interestingly, the free **Mistral 2 Large** is much more successful in this setup — it encoded the whole play:

https://chat.mistral.ai/chat/5b22b8a1-917a-4faa-be7d-d51ea3e32642

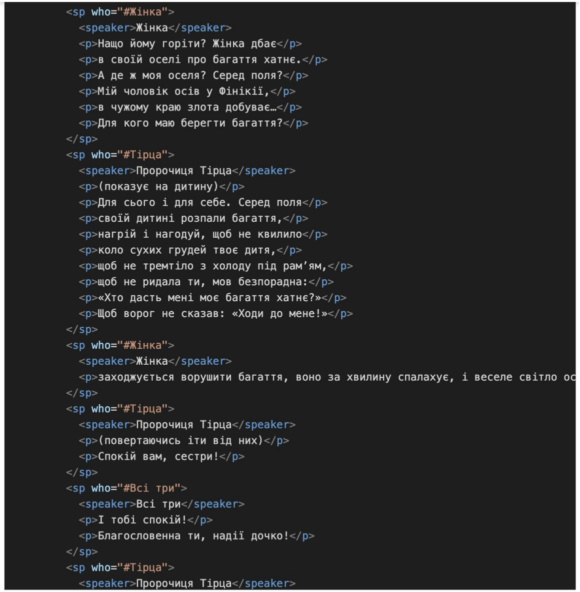


```xml
    <sp who="#Жінка">
      <speaker>Жінка</speaker>
      <p>І тобі спокій!</p>
      <p>Спокій нам всім… спокій і смерть.</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>Чого?</p>
      <p>Живий наш бог, і ми живі.</p>
    </sp>
    <sp who="#Жінка">
      <speaker>Жінка</speaker>
      <p>В руїнах!</p>
      <p>(Притуляється лицем до дитини і плаче. Дитина квилить крізь сон.)</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>Не напувай слізьми малого сина,</p>
      <p>не нагодовуй молоком недолі.</p>
    </sp>
    <sp who="#Дівчина">
      <speaker>Дівчина</speaker>
      <p>Покинутим де іншої взять страви?</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>Чому ти звеш покинутими вас?</p>
    </sp>
    <sp who="#Дівчина">
      <speaker>Дівчина</speaker>
      <p>Мій брат в Ассірії, у Ніневії.</p>
    </sp>
    <sp who="#Дівча-підліток">
      <speaker>Дівча-підліток</speaker>
      <p>Мій батько в Вавілоні другий рік.</p>
    </sp>
    <sp who="#Жінка">
```



```xml
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>Але вас три й дитина межи вами.</p>
      <p>Той не покинутий, хто сам не кинув.</p>
      <p>Паліть багаття, поки встане день,</p>
      <p>щоб вам було видніше при роботі.</p>
    </sp>
    <sp who="#Дівчина">
      <speaker>Дівчина</speaker>
      <p>(кладе кужілку долі)</p>
      <p>Робота наша марна. Хто ж то буде</p>
      <p>носити шати сі, як напряду?</p>
      <p>Мій брат в Ассірії, в ворожім війську,</p>
      <p>неволячи, в неволі сам загине.</p>
      <p>На кого дбати?</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>Дбай сама на себе,</p>
      <p>то будеш вільна, — перед ворогами</p>
      <p>не скажеш: «Хто мені одежу дасть?»</p>
      <p>Не скаже ворог: «Я тебе вдягаю».<p>
    </sp>
    <sp who="#Дівчина">
      <speaker>Дівчина</speaker>
      <p>здіймає кужілку й знову береться прясти.</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>(до дівчини-підлітка, що, зажурившись, склала над роботою руки)</p>
      <p>Чого сидиш, хіба скінчила працю?</p>
    </sp>
    <sp who="#Дівча-підліток">
      <speaker>Дівча-підліток</speaker>
      <p>Та що кінчати? Подруги плетуть</p>
      <p>сі кошики, щоби носити землю</p>
      <p>на могилки над трупами батьків,</p>
```

But stages are
all speeches



```xml
    <sp who="#Жінка">
      <speaker>Жінка</speaker>
      <p>Нащо йому горіти? Жінка дбає</p>
      <p>в своїй оселі про багаття хатнє.</p>
      <p>А де ж моя оселя? Серед поля?</p>
      <p>Мій чоловік осів у Фінікії,</p>
      <p>в чужому краю злота добуває…</p>
      <p>Для кого маю берегти багаття?</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>(показує на дитину)</p>
      <p>Для сього і для себе. Серед поля</p>
      <p>своїй дитині розпали багаття,</p>
      <p>нагрій і нагодуй, щоб не квилило</p>
      <p>коло сухих грудей твоє дитя,</p>
      <p>щоб не тремтіло з холоду під рам’ям,</p>
      <p>щоб не ридала ти, мов безпорадна:</p>
      <p>«Хто дасть мені моє багаття хатнє?»</p>
      <p>Щоб ворог не сказав: «Ходи до мене!»</p>
    </sp>
    <sp who="#Жінка">
      <speaker>Жінка</speaker>
      <p>заходжується ворушити багаття, воно за хвилину спалахує, і веселе світло ос
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
      <p>(повертаючись іти від них)</p>
      <p>Спокій вам, сестри!</p>
    </sp>
    <sp who="#Всі три">
      <speaker>Всі три</speaker>
      <p>І тобі спокій!</p>
      <p>Благословенна ти, надії дочко!</p>
    </sp>
    <sp who="#Тірца">
      <speaker>Пророчиця Тірца</speaker>
```

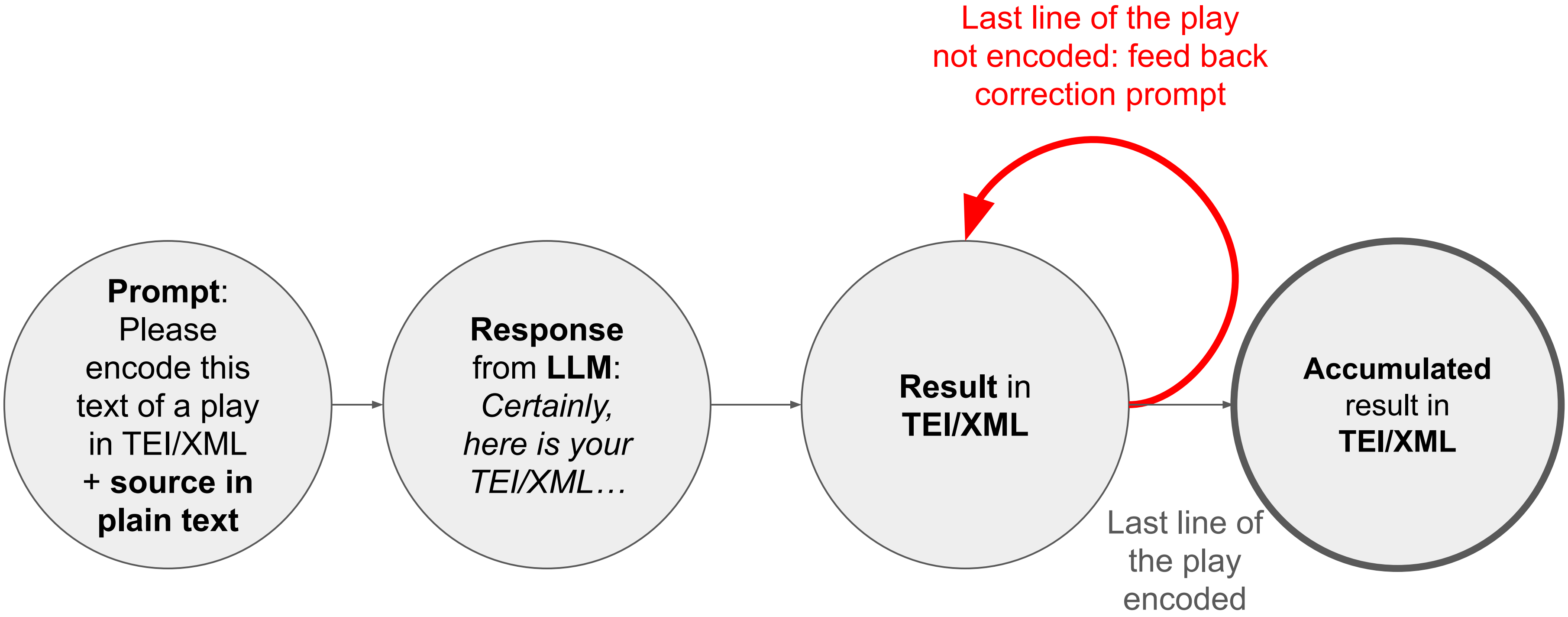
and
so
on
…

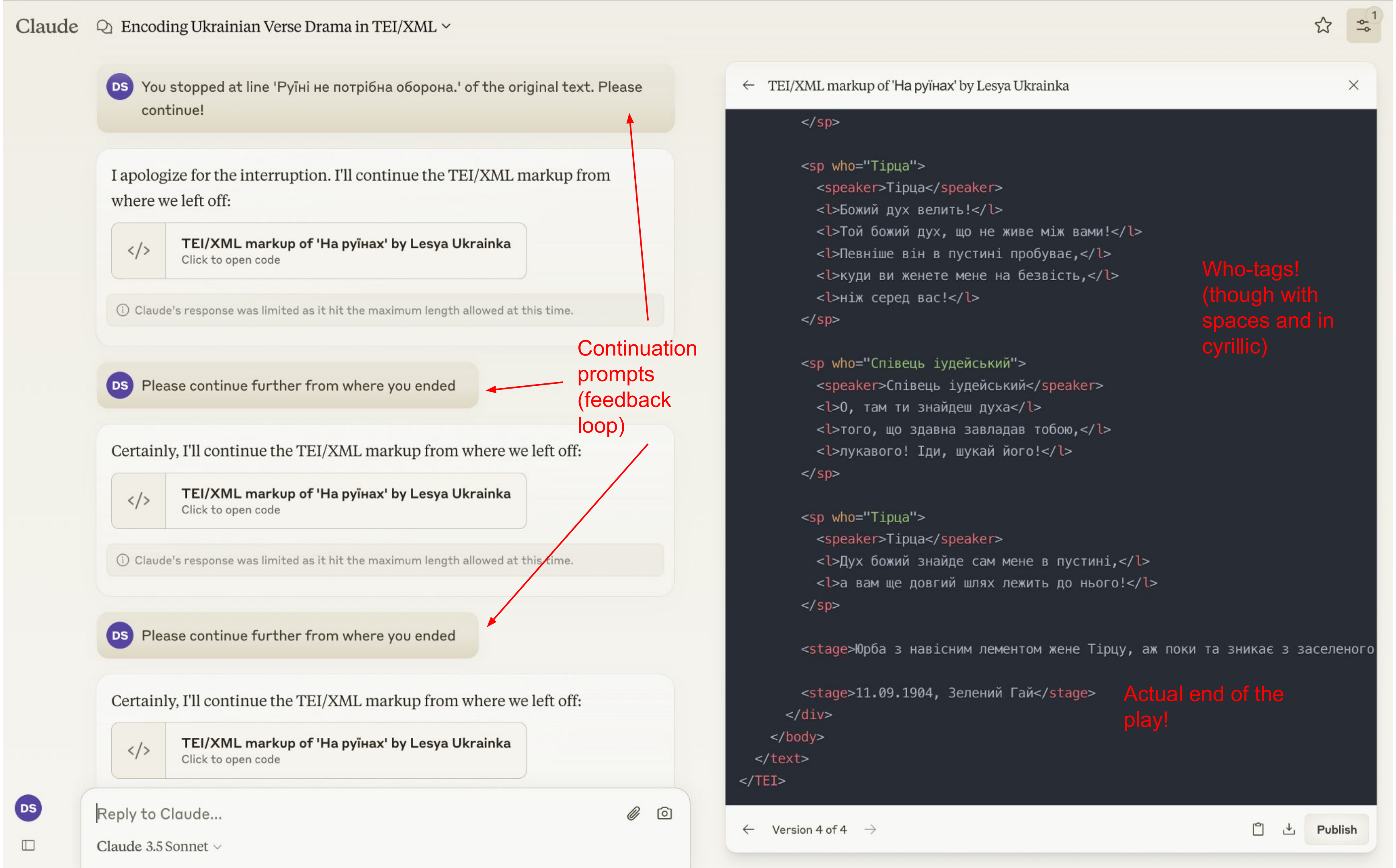
Obvious solution: use the *dialogic* functionality of modern LLMs
(the *Chat* of ChatGPT)

Scenario 2: **multi-step** zero-shot with a **feedback loop**

1. User asks a slightly **more specific initial prompt**: «Please produce a TEI/XML markup of the following Ukrainian play following the TEI/XML standards for drama. This play is in {verse|prose}, so lines spoken by characters should be encoded with {<l>|<p>}. Do not leave or omit any of the text of the original. Do not invent any new text. Do not stop until the whole play is processed»
2. Model responds affirmative and prompts the user to provide text
3. User provides text of the play
4. Model gives output
5. User gives **correction prompt**: You've skipped some lines. Do not skip any lines! Do not add XML comments saying that the dialogue continues! Encode each line as TEI tags instead. This is a very important task for an international research project.
6. Model gives updated output
7. User **asks model to continue** several times (**repeated feedback loop**)

Prompting scenario 2: **multi-step** zero-shot with a **feedback loop** — as a finite-state automaton:





DS You stopped at line 'Руїні не потрібна оборона.' of the original text. Please continue!

I apologize for the interruption. I'll continue the TEI/XML markup from where we left off:

Claude's response was limited as it hit the maximum length allowed at this time.

DS Please continue further from where you ended

Certainly, I'll continue the TEI/XML markup from where we left off:

Claude's response was limited as it hit the maximum length allowed at this time.

DS Please continue further from where you ended

Certainly, I'll continue the TEI/XML markup from where we left off:

Continuation prompts (feedback loop)

TEI/XML markup of 'На руїнах' by Lesya Ukrainka

```xml
        </sp>

        <sp who="Тірца">
          <speaker>Тірца</speaker>
          <l>Божий дух велить!</l>
          <l>Той божий дух, що не живе між вами!</l>
          <l>Певніше він в пустині пробуває,</l>
          <l>куди ви женете мене на безвість,</l>
          <l>ніж серед вас!</l>
        </sp>

        <sp who="Співець іудейський">
          <speaker>Співець іудейський</speaker>
          <l>О, там ти знайдеш духа</l>
          <l>того, що здавна завладав тобою,</l>
          <l>лукавого! Іди, шукай його!</l>
        </sp>

        <sp who="Тірца">
          <speaker>Тірца</speaker>
          <l>Дух божий знайде сам мене в пустині,</l>
          <l>а вам ще довгий шлях лежить до нього!</l>
        </sp>

        <stage>Юрба з навісним лементом жене Тірцу, аж поки та зникає з заселеного

        <stage>11.09.1904, Зелений Гай</stage>
      </div>
    </body>
  </text>
</TEI>
```

Who-tags! (though with spaces and in cyrillic)

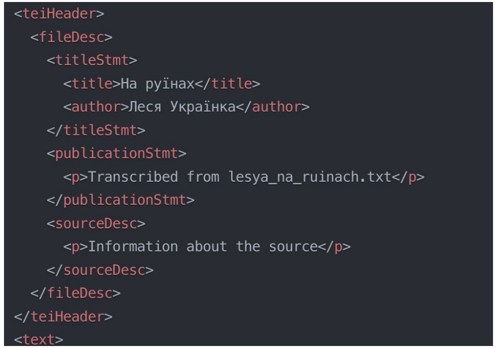
Actual end of the play!

Scenario 2 examples:

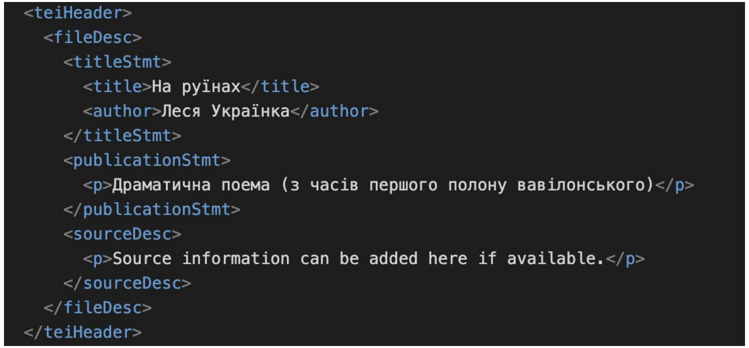
- Claude required 3 feedback runs; here new approach made even the first output before feedback much better (roughly **25% of the play was already there from the first prompt**) and also **added the ‘who’ attributes** which were not there before: https://claude.site/artifacts/ae2def61-8f46-4f87-a757-7c159a85b5ec
- Mistral did it in one go as before, and also learned to **add ‘who’ attributes** (in English)
- ChatGPT required 6 feedback rounds to encode the entire play (with small omissions), but **lost the who attributes:** https://chatgpt.com/share/6707f22c-4e90-800c-9fc5-a2b13ca67a7a
- When assembled correctly from chunks, the files are **well-formed and TEI-all valid**

All three examples still lacking the metadata on characters:

```xml
<teiHeader>
    <fileDesc>
        <titleStmt>
            <title>На руїнах</title>
            <author>Леся Українка</author>
            <respStmt>
                <resp>Edited and encoded by</resp>
                <name>TEI XML Editor</name>
            </respStmt>
        </titleStmt>
        <publicationStmt>
            <p>Published in TEI XML format</p>
        </publicationStmt>
        <sourceDesc>
            <p>Original text provided by user</p>
        </sourceDesc>
    </fileDesc>
</teiHeader>
```



```xml
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>На руїнах</title>
      <author>Леся Українка</author>
    </titleStmt>
    <publicationStmt>
      <p>Transcribed from lesya_na_ruinach.txt</p>
    </publicationStmt>
    <sourceDesc>
      <p>Information about the source</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
<text>
```

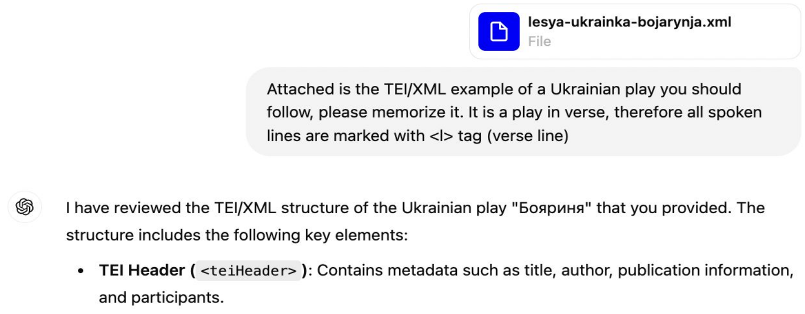


```xml
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>На руїнах</title>
      <author>Леся Українка</author>
    </titleStmt>
    <publicationStmt>
      <p>Драматична поема (з часів першого полону вавілонського)</p>
    </publicationStmt>
    <sourceDesc>
      <p>Source information can be added here if available.</p>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

Can we make them all better with a bit of few-shot learning on existing DraCor markup?

Scenario 3: **multi-step** with a feedback loop and **initial example**

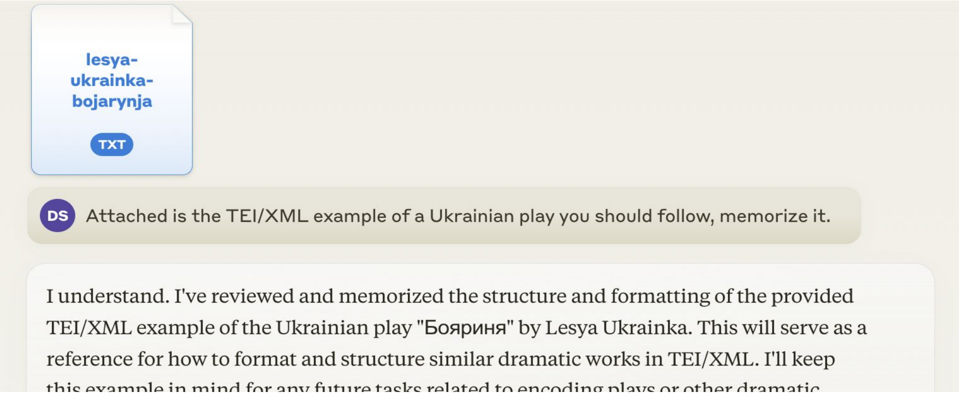
1. User **gives the model an existing DraCor TEI/XML** with a play in the same language and the same mode (verse/prose) and asks to memorize it
2. Then all the steps from Scenario 2



Attached is the TEI/XML example of a Ukrainian play you should follow, please memorize it. It is a play in verse, therefore all spoken lines are marked with <l> tag (verse line)

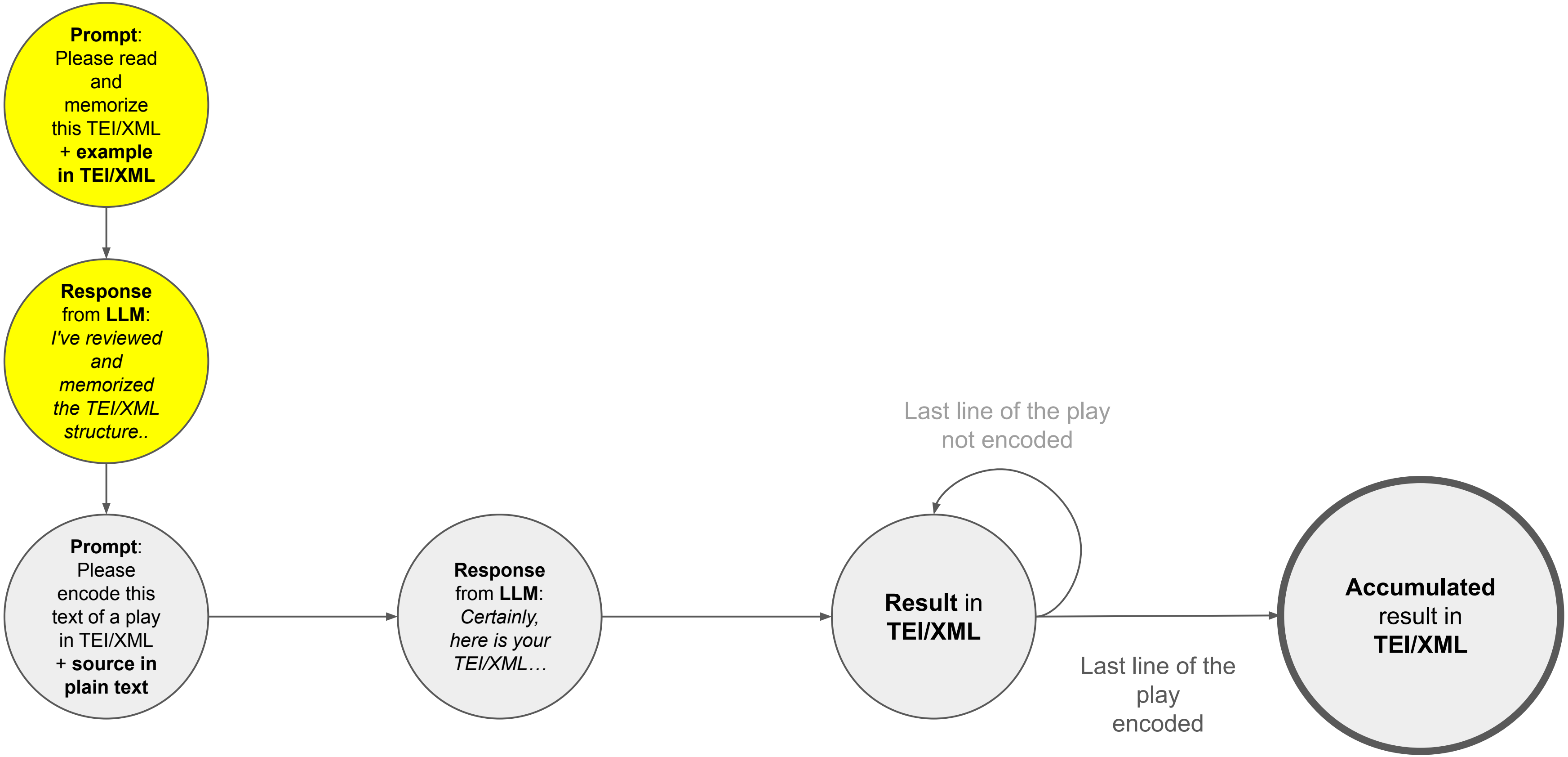
I have reviewed the TEI/XML structure of the Ukrainian play "Бояриня" that you provided. The structure includes the following key elements:

- **TEI Header (** `<teiHeader>` **)**: Contains metadata such as title, author, publication information, and participants.



Attached is the TEI/XML example of a Ukrainian play you should follow, memorize it.

I understand. I've reviewed and memorized the structure and formatting of the provided TEI/XML example of the Ukrainian play "Боярня" by Lesya Ukrainka. This will serve as a reference for how to format and structure similar dramatic works in TEI/XML. I'll keep this example in mind for any future tasks related to encoding plays or other dramatic



Prompt: Please read and memorize this TEI/XML + **example in TEI/XML**

Response from **LLM**: *I've reviewed and memorized the TEI/XML structure..*

Prompt: Please encode this text of a play in TEI/XML + **source in plain text**

Response from **LLM**: *Certainly, here is your TEI/XML…*

Result in **TEI/XML**

Last line of the play not encoded

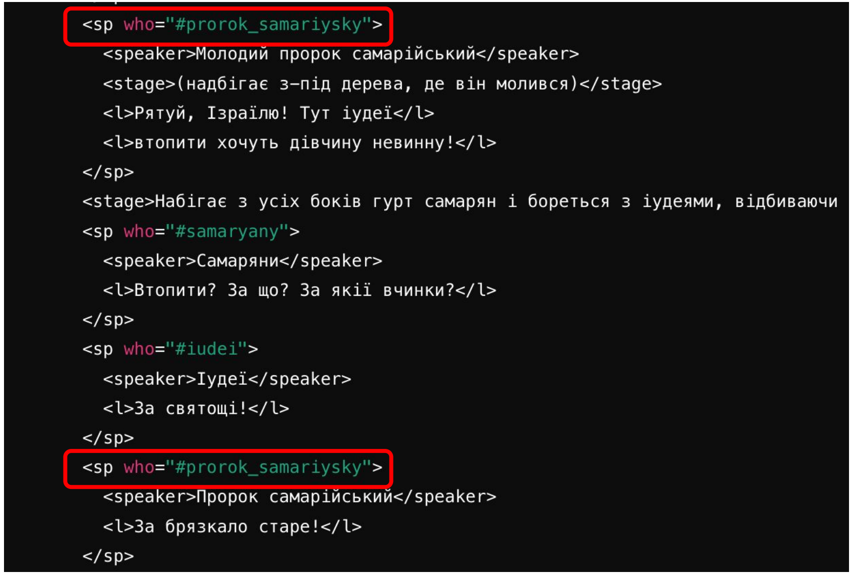
Last line of the play encoded

Accumulated result in **TEI/XML**

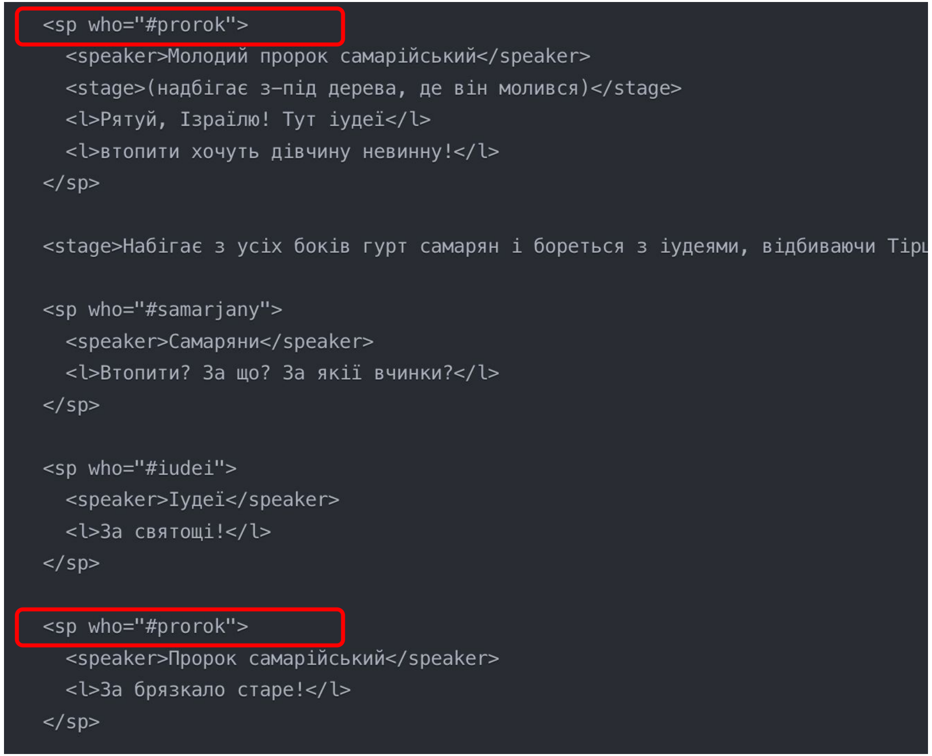
Scenario 3 examples:

- Mistral refused to work in that setting (at least in those couple of times that we tried) — both the example and the new play together were too much for it
- Further improvements for ChatGPT, which now has character IDs — and they even look totally dracor-ish!: https://chatgpt.com/share/91b90ef2-1067-41c5-95b8-13ebeed394a8
- Both Claude and ChatGPT learned to identify characters by ID even if the <speaker> text differs a bit

Identified characters (this was never asked explicitly)



```
<sp who="#prorok_samariysky">
  <speaker>Молодий пророк самарійський</speaker>
  <stage>(надбігає з-під дерева, де він молився)</stage>
  <l>Рятуй, Ізраїлю! Тут іудеї</l>
  <l>втопити хочуть дівчину невинну!</l>
</sp>
<stage>Набігає з усіх боків гурт самарян і бореться з іудеями, відбиваючи
<sp who="#samaryany">
  <speaker>Самаряни</speaker>
  <l>Втопити? За що? За якії вчинки?</l>
</sp>
<sp who="#iudei">
  <speaker>Іудеї</speaker>
  <l>За святощі!</l>
</sp>
<sp who="#prorok_samariysky">
  <speaker>Пророк самарійський</speaker>
  <l>За брязкало старе!</l>
</sp>
```



```
<sp who="#prorok">
  <speaker>Молодий пророк самарійський</speaker>
  <stage>(надбігає з-під дерева, де він молився)</stage>
  <l>Рятуй, Ізраїлю! Тут іудеї</l>
  <l>втопити хочуть дівчину невинну!</l>
</sp>

<stage>Набігає з усіх боків гурт самарян і бореться з іудеями, відбиваючи Тір

<sp who="#samarjany">
  <speaker>Самаряни</speaker>
  <l>Втопити? За що? За якії вчинки?</l>
</sp>

<sp who="#iudei">
  <speaker>Іудеї</speaker>
  <l>За святощі!</l>
</sp>

<sp who="#prorok">
  <speaker>Пророк самарійський</speaker>
  <l>За брязкало старе!</l>
</sp>
```

Metadata improvements

```xml
<TEI xml:lang="ukr" xmlns="http://www.tei-c.org/ns/1.0">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">На руїнах</title>
        <title>Драматична поема</title>
        <author>
          <persName>
            <forename>Леся</forename>
            <surname>Українка</surname>
          </persName>
          <persName xml:lang="eng">
            <forename>Lesya</forename>
            <surname>Ukrainka</surname>
          </persName>
        </author>
      </titleStmt>
      <publicationStmt>
        <publisher xml:id="ukrdrama">Ukrainian Drama Collection</publisher>
        <availability>
          <licence>
            <ab>Public Domain</ab>
          </licence>
        </availability>
      </publicationStmt>
      <sourceDesc>
        <bibl type="originalSource">
          <title>На руїнах</title>
          <author>Леся Українка</author>
          <pubPlace>Зелений Гай</pu  ce>
          <date>11.09.1904</date>
```

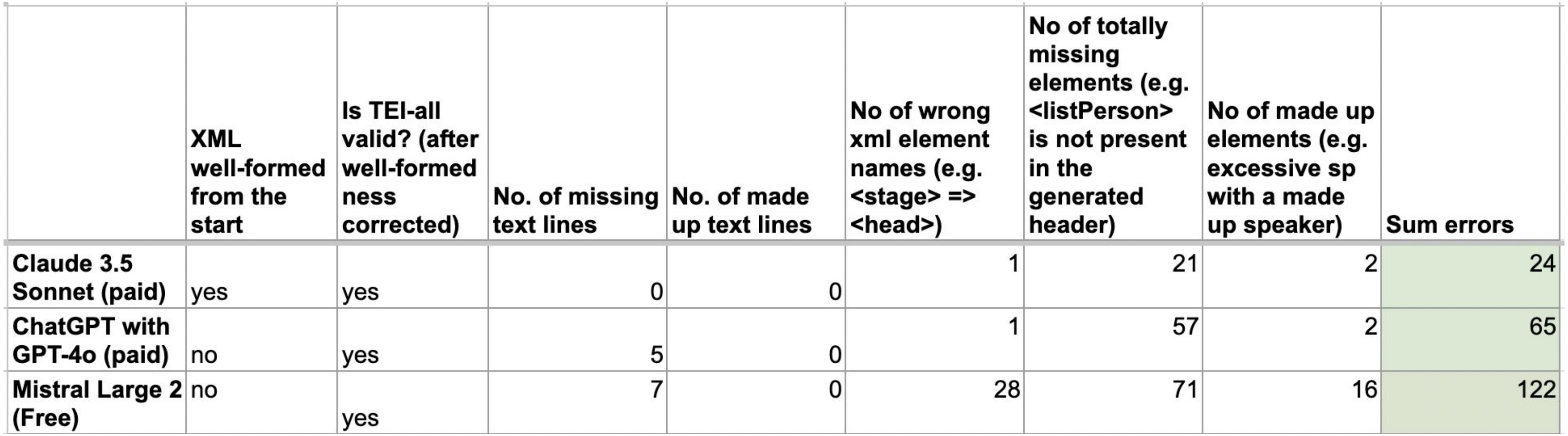
```xml
<profileDesc>
  <particDesc>
    <listPerson>
      <person sex="FEMALE" xml:id="tirca">
        <persName>Тірца</persName>
      </person>
      <person sex="FEMALE" xml:id="zhinka">
        <persName>Жінка</persName>
      </person>
      <person sex="FEMALE" xml:id="divchyna">
        <persName>Дівчина</persName>
      </person>
      <person sex="FEMALE" xml:id="divcha_pidlitok">
        <persName>Дівча-підліток</persName>
      </person>
      <person sex="MALE" xml:id="cholovik">
        <persName>Чоловік</persName>
      </person>
      <person sex="MALE" xml:id="staryj">
        <persName>Старий</persName>
      </person>
      <person sex="MALE" xml:id="spivets">
        <persName>Співець</persName>
      </person>
      <person sex="MALE" xml:id="rybalka">
        <persName>Рибалка</persName>
      </person>
      <person sex="MALE" xml:id="prorok_samariyskyi">
        <persName>Пророк самарійський</persName>
      </person>
      <person sex="MALE" xml:id  evit">
        <persName>Левіт</persName>
      </person>
      <personGrp sex="MALE" xml:id="samaryany">
        <name>Самаряни</name>
```

```xml
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:id="u000005" xml:lang="ukr">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title type="main">На руїнах</title>
        <author>
          <persName>
            <forename>Леся</forename>
            <surname>Українка</surname>
          </persName>
          <idno type="wikidata">Q298033</idno>
        </author>
      </titleStmt>
      <publicationStmt>
        <publisher xml:id="dracor">DraCor</publisher>
        <idno type="URL">https://dracor.org</idno>
        <availability>
          <licence>
            <ab>CC0 1.0</ab>
            <ref target="https://creativecommons.org/publicdomain/zero/1.0/"
          </licence>
        </availability>
      </publicationStmt>
      <sourceDesc>
        <bibl type="digitalSource">
          <name>Wikisource</name>
          <idno type="URL">https://uk.wikisource.org/wiki/На_руїнах_(Леся_Ук
        </bibl>
      </sourceDesc>
    </fileDesc>
    <profileDesc>
```

```xml
<profileDesc>
  <particDesc>
    <listPerson>
      <person xml:id="tirca" sex="FEMALE">
        <persName>Тірца</persName>
      </person>
      <person xml:id="zhinka" sex="FEMALE">
        <persName>Жінка</persName>
      </person>
      <person xml:id="divchyna" sex="FEMALE">
        <persName>Дівчина</persName>
      </person>
      <person xml:id="divcha" sex="FEMALE">
        <persName>Дівча-підліток</persName>
      </person>
      <person xml:id="cholovik" sex="MALE">
        <persName>Чоловік</persName>
      </person>
      <person xml:id="staryi" sex="MALE">
        <persName>Старий</persName>
      </person>
      <person xml:id="spivec" sex="MALE">
        <persName>Співець</persName>
      </person>
      <person xml:id="prorok" sex="MALE">
        <persName>Пророк самарійський</persName>
      </person>
      <person xml:id="levit" sex="MALE">
        <persName>Левіт</persName>
      </person>
    </listPerson>
  </particDesc>
```

Attempt at quality assessment (best version of each model)

- Script that counts the missing lines
- Script that counts the 'made-up' lines (with some manual post-assessment)
- Semi-manual tag structure comparison



	XML well-formed from the start	Is TEI-all valid? (after well-formed ness corrected)	No. of missing text lines	No. of made up text lines	No of wrong xml element names (e.g. <stage> => <head>)	No of totally missing elements (e.g. <listPerson> is not present in the generated header)	No of made up elements (e.g. excessive sp with a made up speaker)	Sum errors
Claude 3.5 Sonnet (paid)	yes	yes	0	0	1	21	2	24
ChatGPT with GPT-4o (paid)	no	yes	5	0	1	57	2	65
Mistral Large 2 (Free)	no	yes	7	0	28	71	16	122

Takeaways

- With enough input on the desired TEI format, **LLMs can produce almost ready-made DraCor TEI/XMLs** (if you poke them enough times)
- **Different strategies work better with different LLMs**, e.g. in our case:
    - Mistral was the best ‘from-scratch’ model
    - Claude had huge improvement at stage 2, was already almost perfect
    - ChatGPT did its best after looking at an example at stage 3
- Results are **not very stable**, e.g. with ChatGPT currently it became somewhat harder to reproduce scenario 3 as smoothly as it worked earlier
- **Length** generally **remains a challenge** (which is understandable given the nature of the Transformer architecture)

2025:
current perspectives

Status quo

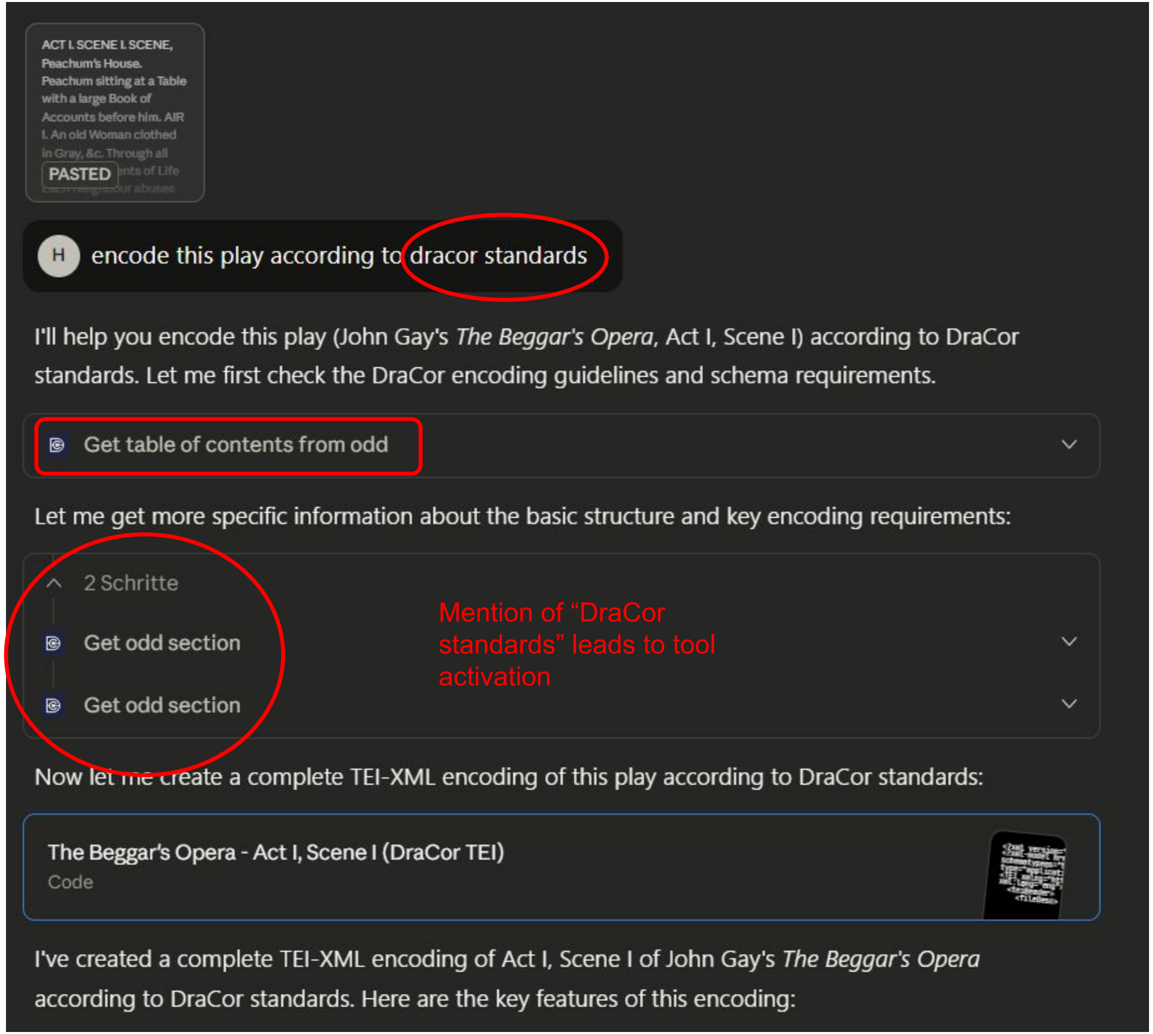
- To our knowledge the most recent additions to DraCor (RoDraCor, IbsDraCor, etc.) have not been encoded with the help of LLMs…
- … and many contributors (e.g. for ArDraCor) still prefer *easydrama* for its reliability as compared to LLMs..
- Elsewhere, scholars are developing evaluation framework for assessing large language models in text encoding tasks (e.g. Strutz & Vogeler 2025; see also work by Christopher Pollin and others) → we need one for drama as well.

What about MCP integration?

Seems really promising, but much more experimenting needs to be done.

Next steps:

- evaluating how zero-shot MCP-supported encoding performs compared to “dialogic” encoding
- (long term) prototyping Interface for direct LLM-powered play encoding within the DraCor website



ACT I. SCENE I. SCENE, Peachum's House. Peachum sitting at a Table with a large Book of Accounts before him. AIR I. An old Woman clothed in Gray, &c. Through all

PASTED

H encode this play according to dracor standards

I'll help you encode this play (John Gay's *The Beggar's Opera*, Act I, Scene I) according to DraCor standards. Let me first check the DraCor encoding guidelines and schema requirements.

Get table of contents from odd

Let me get more specific information about the basic structure and key encoding requirements:

2 Schritte

Get odd section

Get odd section

Now let me create a complete TEI-XML encoding of this play according to DraCor standards:

The Beggar's Opera - Act I, Scene I (DraCor TEI)
Code

I've created a complete TEI-XML encoding of Act I, Scene I of John Gay's *The Beggar's Opera* according to DraCor standards. Here are the key features of this encoding:

Mention of “DraCor standards” leads to tool activation

Thank you!

skorinkin@uni-potsdam.de
giovannini@uni-potsdam.de

Related abstract:

Giovannini, Luca, and Daniil Skorinkin (2024). "Towards a LLM-powered encoding workflow for plays / Hacia un flujo de trabajo de codificación para obras de teatro impulsado por LLM". In: *TEI2024 Book of Abstracts*. Universidad del Salvador, 2024, pp. 94-95.